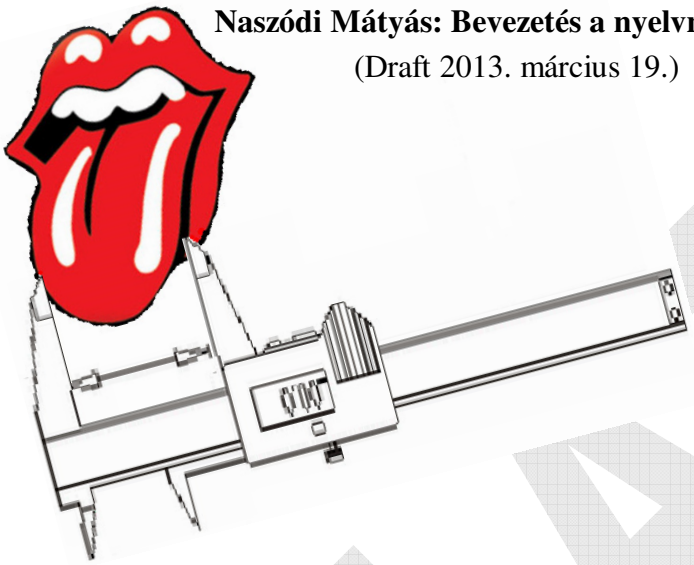
Naszódi Mátyás: Bevezetés a nyelvmérnökségbe

(Draft 2013. március 19.)



Előszó

A jegyzet a hasoncímű egyetemi előadásom anyaga. A témát a 80-as évek végén kezdtük el kedves kollégámmal, Farkas Ernővel oktatni. Ő már rég nincs köztünk, de nélküle nem hiszem, hogy képes lettem volna ebbe témába elmélyülni.

Akkoriban a nyelvmérnökség (language engineering) még ismeretlen fogalom volt. Nem is így neveztük. Ma inkább nyelvtechnológiának hívják. Én maradtam a mérnöki jelzőnél, mert a kezdetekben is azt vallottuk, ma is az a véleményem, hogy a számítógépes nyelvi modellezésnek az a lényege, hogy használható legyen, ez pedig mérnöki feladat.

Magam matematikusi végzettséggel rendelkezem, de az egyetem alatt és azóta is informatikussághoz közelebb állok. A nyolcvanas évek közepétől foglalkozom a természetes nyelvek – főként a magyar – számítógépes megközelítésével. Sok gyakorlati tapasztalat gyűlt fel bennem, és szeretném átadni ezeket másoknak. Az évek során a nyelvfelfogás a nyelvészeknél, a technológia a számítógépes programozásban sokat változott. Az egyetemi kurzusom anyaga is sokat módosult a kezdetek óta.

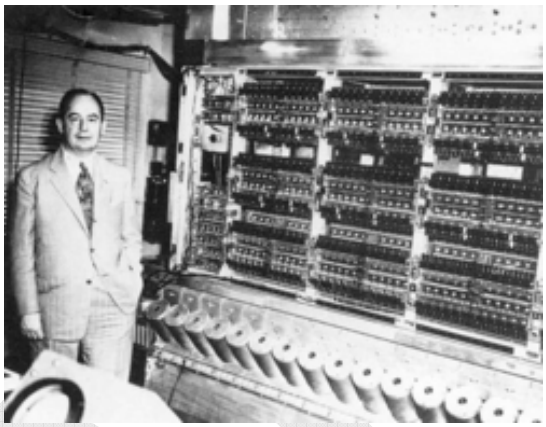
Az anyag egyrészt nyelvészeti, másrészt informatikai meggondolásokat tartalmaz. Nyelvészi képzésben sohasem részesültem, de az anyanyelvét mindenki mesterfokon bírja. Persze, a nyelvészektől sokat tanultam. Informatikusképzés abban az időben, amikor végeztem az egyetemen, nem volt. Magam is úgy lettem azzá, mint korosztályom más matematikus-, fizikus-, villamosmérnök-képzettségű informatikus, önképzéssel. Azóta persze oktattam is informatikusokat, legalábbis speciális matematikai tárgyakra.

Hasonló témát a 80-as években egyáltalán nem oktattak, külföldön sem. Ma már itthon is léteznek egyetemek, ahol a számítógép és a természetes nyelv határtudománya tantárgy. Jó jegyzet magyarul mégsem született. Legtöbbször angol nyelvű elfogadott irodalomból indulnak ki. Én a szokásos ilyen jellegű jegyzetektől eltérek. Egyrészt sok olyan matematikai fejezetet kihagyok, melyet informatikusoknak kötelezően előírnak, másrészt a példáim alapvetően, de nem kizárólag magyar nyelvre vonatkoznak. A témákat is szűkítem. A szótan és a mondattan a két félévi anyagom fő tárgya. Ezekben a munkáim során felfedezett nyelvi összefüggéseket tárom az olvasó, tanuló elé, melyeket, azóta hogy felismertük, tőlünk függetlenül, igazi nyelvészek is megírtak. Csak érintőlegesen foglakozom beszédfelismeréssel, szemantikával, szövegbányászattal. Ezeknek a témáknak nálam sokkal jobb tudorai vannak.

Tartalomjegyzék

1. Bevezetés

1.1 A számítógépes nyelvészet történetéből



A számítógép megjelenésével hamarosan felmerült annak igénye, hogy természetes nyelvek fordítására használják. A világháborút megelőző időkben és alatta elsősorban statisztikák készítésére, tudományos számításokra, és titkosírások feltörésére használták. Az akkori gépek teljesítménye nem is tette lehetővé, hogy összetettebb feladatokra hasznosítsák a komputereket. Bár nagy termeket töltöttek be (eleinte relés logikára épültek, később csöves számítógépek jelentették a technika csúcsát), csak energiaigényük volt nagy. Mi sem jellemzőbb, minthogy egy több ezer szavas gép már nagy teljesítményűnek számított.

A félvezetős technológia hozta meg az első reményt, hogy valóban nyelvi alkalmazásokra használják a számítógépet. Körülbelül erre az időre tehető Noam Chomsky munkájának ismertté válása, amelyben megalapozta e tudományt.

Az, hogy a fordítást, az egyik legfáradtságosabb szellemi munkát automatizálják, szintén a világháborút követő időszakban vált valós igénnyé, a technikai fejlődés „nemzetközösülésével", de nem utolsósorban a hírszerző szervek azon törekvésével, hogy idegen államokból származó szövegeket érthetővé tegyék az alkalmazottai számára. Nem véletlen, hogy az amerikai hadügyminisztérium (D.O.D.) adta a legnagyobb támogatást a kutatásokra.

Mintegy tízéves munka után, 1966-ban Johnson, az akkori elnök összehívta a tojásfejűeket (a szakértőket), hogy tisztázza, mikor várható már, hogy a számítógépet beállítsák a csatasorba. Akkoriban a válasz, az ALPAC jelentés sommás volt:

A számítógépes fordítás
1. sokkal lassabb,
2. sokkal drágább,
3. sokkal rosszabb, mint a humán fordítás

A jelentés a pénzek megvonásához vezetett. A kutatók, fejlesztők egy része átpártolt az akkor rohamos fejlődésnek induló és anyagi jövedelmezőséggel kecsegtető gépi nyelvészethez, míg egy kisebb, megszállott társaság tovább dolgozott egyetemi, akadémiai keretek között. A programnyelvek ekkor hihetetlen fejlődésen mentek át. Elemzési, fordítási technológiák alakultak ki, nyelvek születtek minden jobb egyetemen, és ez a későbbiekben áldásos hatással volt a természetes nyelv feldolgozói számára is.

Csipkerózsika álmukból a technikai fejlődés, illetve az újabb igények ébresztették fel – no nem a kutatókat – hanem a pénzt osztogatókat. Ne feledjük el, hogy a hetvenes években, ha nem is Magyarországgal, de létezett a többnyelvű Európai Közösség, ahol – legalábbis deklaráció szintjén – minden nyelv egyenrangú, ezért minden közös határozat, törvény, rendelet minden nyelven meg kell, hogy jelenjen. Nem beszélek a multinacionális cégekről, a kereskedelmi, kisebbségi törvényekről, ami mind azt jelenti, hogy az amúgy is egyre nagyobb mennyiségben megjelenő írásos anyagok fordítását fizikailag sem lehet tisztán emberi erőre bízni.

Másrészt a félvezető technológiában megjelentek a nagybonyolultságú integrált áramkörök, amelyek lehetővé tették a gépek fizikai képességének rohamos növekedését, úgy sebességben, mint tárkapacitásban.

Így a hetvenes évek során megint lendületet vett a kutatás. Ötödik, hatodik generációs számítógépekről beszéltek Amerikában, Japánban, de nálunk is, és azt hittük, tíz éven belül emberi nyelven vezéreljük a folyamatokat, a gép a mi nyelvünkön válaszol, ne adj Isten, tolmácsként dolgozik. A gépi fordítás terén iskolák alakultak ki, és kecsegtető eredményeket mutattak Moszkvában, az M.I.T.-n Grenoble-ban és másutt. Nagy reményt fűztek a logikai programozás felhasználásához, a Prolog alkalmazásához.

Rengeteg hasznosítható eredmény született. A számítógépet egyre gyakrabban alkalmazták humán célokra, de ipari-kereskedelmi fordító talán csak egy született ebben az időben, a Systran. Mint kiderült, a látványos kezdeti eredmények után az igazit, a FORDÍTÓ-t nem olyan könnyű megteremteni. Ennek egyik oka, hogy az egyszerű szabályokkal leírható jelenségek megadhatják ugyan a fordítás vázát, de azt kitölteni a nyelv apró és számtalan jelenségével nem lehet. Ehhez újabb fejlődésre kellett várni.

Születtek persze gyakorlati eredmények. Ezek – mondhatnám – melléktermékei a fejlesztéseknek, de tömeges felhasználásuk miatt nagy jelentőséggel bírnak. Ilyenek a helyesírás-ellenőrzők, stiláris korrektorok, gépi szóelválasztók, mindenki által hozzáférhető gépi szótárak, szinonimatárak. Ritkábban, de alkalmazzák a tudomány eredményeit szövegkereséseknél, karakterfelismerésnél, témafigyelésnél.

A fordítás ma már nem álom. Egyrészt az igény ma nagyobb, mint valaha, gondoljunk az ezernyelvű internet galaxisra, másrészt a számítógépek kapacitásának folytonos fejlődése meg is teremti a lehetőségét, hogy olyan – a korábbinál nagyobb idő és memóriaigényű – módszereket alkalmazzanak, mely ezelőtt tíz évvel alkalmazhatatlannak tűntek.

Ma már bizonyos nyelvek között létezik gépi vagy géppel segített fordítás. Ha ma kérdezné meg a pénzen ülő főhivatalnok a tojásfejűeket, akkor a válasz már néhány nyelvpárra a következő lenne:

A számítógépes fordítás
1. sokkal gyorsabb,
2. sokkal olcsóbb,
3. és bizonyos esetekben sokkal jobb, mint az emberi fordítás

A magyarországi fejlődésről is had ejtsek pár szót. A keleti blokkban a számítógép-tudomány tiltott diszciplína volt a személyi kultusz idején. Magyarországon, a számítógép hivatalos elfogadása után elég hamar kísérleteket tettek gépi fordításra. Az első demonstratív kísérletek Ábrahám Samu nevéhez fűződnek. A program mintegy húszszavas szótárral, de nyelvtani alapon volt képes fordítani a hatvanas években. Ebbe a munkába folyt bele Kiefer Ferenc, Varga Dénes, Dömölki Bálint és mások. Mindannyian a szakma nagy öregjeinek tekinthetők. A dátumot nem ismerem, de a számítógép alkalmazásában talán megelőzte őket Papp Ferenc, akinek vezetésével (akkor Debrecenben) lyukkártyás rendszer segítségével készítették el a magyar atergo (szóvégmutató) szótárat. A hetvenes évekből kiemelném Füredi Mihályt, aki elkészítette a magyar nyelv gyakorisági szótárát.

A lanyhának mutatkozó magyarországi fejlődés talán a nyolcvanas években volt eredményes. Ekkor jelentek meg Prószéky Gábor munkái a magyar morfológia „algoritmizálásáról", a SzTAKI-ban Bach Iván vezetésével munkák magyar mondatok elemzéséről, Tihanyi László és társainak AMI/NMI számítógépes szótárai, a BME-n beszédszintetizátorok, és más, esetleg külföldön tanító/dolgozó kutató magyar nyelvre vonatkozó munkái. Az első tömegalkalmazások a 90-es évek elején megjelenő szpellerek és elválasztók voltak. Az ezredfordulóig a kutatások soha nem kaptak kiemelt (hazai) támogatást, ennek ellenére az eredmények bíztatóak voltak.

Ma az egyetemeken kívül a Nyelvtudományi Kutatató Intézetben foglalkoznak a témával, illetve kutató-fejlesztő, de profitorientált cég egy van, a MorphoLogic, a melyik magára vállalta a kutatás és az ipar közötti űr betöltését – célja, hogy hasznosítsa a gyakorlatban mindazt, ami a nyelvi mérnökségből a magyar nyelv terén hasznosítani kell és lehet.

A fent említett helyesírás-ellenőrzőn kívül a legkülönbözőbb szótárak és fordítástámogató rendszerek léteznek, szövegindexelő, kereső rendszerek, optikai felismerők alkalmazzák a nyelvi tudást. Az idő folyamán már megjelentek a használható, bár tökéletlen gépi eszközök, melyek segítik a fordítást. A tömeges gépi fordítás az európai csatlakozással nélkülözhetetlen válik. Ismervén a fejlődést, hamarosan várható a magyar nyelvű beszédfelismerő is. Hogy a hamar mit jelent, egy évet vagy ötöt, nem tudom, de e könyvet azért írtam, hogy többen segítsék e fejlődést.

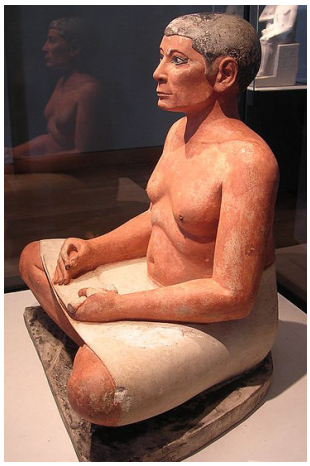
1.2 A számítógépes nyelvészet résztudományai és alkalmazásai

A számítógépes nyelvészet osztályozásának több szempontja lehet. Az egyik, hogy kutatás vagy alkalmazás a célja, a másik, hogy a nyelvészet melyik szintjét célozz, használja. A határok nem élesek. Maga a kutatás is lehet egy alkalmazási terület. A nyelv szintje is „áthatnak". Ezzel a megjegyzéssel értelmezendő a következő táblázat mely minta jellegűen mutat néhány alapvető témát, alkalmazást a következő beosztás szerint.

	kutatás	alkalmazás
karakterek fonémák	fonetikai elemzők, generátorok karakterstatisztikák	karakterkonverterek optikai karakterfelismerők beszédfelismerők, -generátorok
szavak	nyelvészeti szótani rendszerek szótani elemzők szógenerátorok lemmatizálók szótani „cédulázási” munkák támogatása szóstatisztikák, gyakorisági szótárak szótárkészítés támogatása	helyesírás-ellenőrzők elválasztó programok automatikus korrektorok (pl. beszéd-, karakterfelismeréshez) szövegindexelők keresőprogramok elektronikus szótárak
mondatok	nyelvészeti szintaktikai modellek mondatelemzés mondatgenerálás szöveg-egyértelműsítés	nyelvtani és stiláris ellenőrzők gépi és géppel segített fordítás nyelvi ellenőrzés (pl. beszéd-, karakterfelismeréshez) helyes intonáció előállítása (beszédszintetizáláshoz)
szöveg	hiány feloldása szövegstatisztikák korpusznyelvészet	szövegkivonatoló rendszerek szövegszűrők szövegszinkronizációk természetes nyelvű gép-ember kapcsolat

A táblázat nem lehet teljes. Az emberi leleményesség újabb és újabb témákat és alkalmazásokat teremt, talál.

1.3 Az írásbeliség történetéből



Igazán nem tudni, mikor és hogyan alakult ki az írás. Az biztos, hogy az ősember barlangrajza alapvetően nem nevezhető írásnak, mert – bár akár egy történetet is lerajzolhattak képregény formájában – az ábrázolás nem a nyelvhez kötődött. Írásról akkor beszélünk, ha az a nyelv képi mása. Persze mi a nyelv? A nyelv a gondolatok közlésének eszköze, melyben jelek egymásutánjával *linearizált* gondolatot adnak át az emberek egymásnak. Itt lényeges tulajdonság a gondolat linearizálása, mert gondolkozni lehet párhuzamosan is, de beszélt, írott nyelven közölni csak sorosan lehet. A nyelvek mindegyike jelek egymásutánjából épül fel.

Másik jellegzetesség, hogy beszédben, írásban nem magát a dolgokat, hanem annak nyelvi reprezentációját használjuk. A betűírásnál ezt természetesnek vesszük, de pl. az ókori egyiptomiak hieroglifái képekkel jelölték a szavakat. A képek pedig többségükben azt az objektumot ábrázolták, amiről éppen szó volt. Az ilyen írást (szóírást) ideografikus írásmódnak nevezzük. Ma is létezik ideografikus írásmód. A kínaiak mind a mai napig szóírást használnak. A kínai írott szövegekben már gyakorlatilag nem ismerhető fel a szó tartalmának képe, de az ókori hieroglifák sem mindig azt jelentették, amit közvetlen ábrázolt a kép. Ha nem így lenne, akkor az egyiptológusoknak nem lenne akkora teher megfejteni a két-háromezer éves feliratokat. Az ideografikus írás egy nagy hátránya, hogy az írás és a kiejtés elszakad.

Az eddig feltárt legősibb írásos emlékek a babilóniai Uruk város romjai között talált agyagtáblák. Az időszámításunk előtt körülbelül 3000-ben írt feljegyzések – a mintegy 700 jel megfejtése után kiderült – gazdasági jellegűek: gazdasági ügyletek nyilvántartása, állatok, állati termékek jegyzékei, számadatok. Nyilvánvaló, hogy az írás maga ennél régebbi keletű, de a fába és más élő eredetű anyagba rótt jelek a semmibe vesztek az idők során. A korai írások mind képszerűek voltak, de néhány évszázaddal későbbiek már a fonetizálás jegyeit mutatják az írásos emlékek. A fonetikus jelleg (többnyire szótagírás) akkor erősödik meg, amikor eltérő nyelveket beszélő népek kezdték használni. Nagyjából ugyanabban az időben jelentek meg – talán a babilóniai ötlet hatására – más nyelvű népeknél más írásrendszerek. (Pl. az egyiptomi hieroglifák.)

A régészeti leletek alapján i.e. XIII. században a föníciaiak kezdték használni azt az írásmódot, amely ma is legjobban elterjedt és leghatékonyabbnak bizonyult: a betűírást. A betűírás jelei is képekből származnak. Közismert, hogy a mi A (illetve a) betűnk a görög alfa (A illetve α) betűből származik, ami régi (föníciai,

illetve sumér) 𒀭 jelből származik, mely szemlátomást egy ökröt (alfát) reprezentál. A jelnek már csak annyi köze van az ökörhöz, hogy a szó első betűjét jelzi.

A képírás és betűírás között egy nagy különbség, a különböző jelek száma. Míg az ideografikus írásnál minden fogalomnak külön jele van, addig a betűírásnál alapvetően a hangoknak van jele. Az előzőnél több ezer jelet kell megtanulnia az iskolásnak (de több száz kell a minimális írástudáshoz) addig az utóbbinál csak húsz-harmincat. A betűírásnak is több ezer éves múltja van. Még a sumérokat megelőző kultúrákból maradtak fenn olyan nyomok, melyeknél pásztornépek gyöngysorba fűzve hordozták a terelt nyáj egyedeinek mását. Ebben nem csak a gondozott jószágok szerepeltek, hanem a segítő juhászkutya is. Ennek a kutyának (kutyafejnek lelapuló fülekkel) mását lehet fellelni egyes kezdeti írásbeliségekben.

A sumér írás a fonetikán alapul. A hangok helyett, szótagokat ábrázoltak a jelek. Ha igaz az a nyelvre, hogy minden szó mássalhangzóval kezdődik, magánhangzóval végződik, és váltogatják benne egymást a magánhangzók és mássalhangzók, akkor egyszerű kis táblázatba lehet foglalni az írásjeleket. Mondjuk, tizenkét mássalhangzó és nyolc magánhangzó esetén egy 12×8 méretű mátrixba elhelyezhető az összes jel. Így is tették, így az ő ábécéjük nem soros, hanem kétdimenziós volt. Ilyen jellegű írás több is létezett az ókorban, de ma is van szótagíró nép. Például a koreaiak ma is így írnak, a japán katakana jelek is szótagokat jelölnek. Részben ilyennek mondhatók a szemita (arab, héber) nyelvek írása is. Ezeken a nyelveken a (hangsúlytalan) magánhangzók jelölése nem kötelező, pontosabban normális esetben nem jelölik. Akkor szokták kiírni, ha a szöveg a magánhangzók hiányában félreérthető. A mohamedánok szent könyvében, a Koránban kötelező a magánhangzók jelölése, nehogy máshogy értelmezze a hívő a szentírást, de máskülönben nem szokás használni a magánhangzókat. Ebből eredően – mivel írásban hiányosan rögzültek egyes szavak, máshogy beszéli a nyelvet egy algériai arab, mint egy szaúdi. A katolikusok bibliájának is sok értelmezése van, hisz egyes részeit héber közvetítéssel vették át, és ez is lehetőséget teremtett egyes szavak félreértelmezésére.

A betűírás korszerűnek mondható, mert aránylag kevés jellel lehet mindent papírra vetni. A betűírás többnyire a fonetikát követi. Ez akkor is igaz, ha egy hangnak több jel fele meg, illetve ha az írott szöveg (a szó jelentésétől függően) többféleképpen ejthető ki. (pl. angolban).

Az írásbeliség kihat a beszélt nyelvre is. Ha valamit írásban nem jelölünk, akkor a fejlődés során annak kiejtése változhat. Azoknál a nyelveknél, ahol a magánhangzót nem jelölték, erősebb átalakulások mentek végbe a beszélt nyelvben, mint azoknál, ahol az írás tükrözi a beszéd teljes fonetikáját. Az ideografikus írásnak nincs sok köze nincs a kiejtéshez. Bár a kínai írásban a szavak jelei is tartalmaznak fonetikai elemeket, ezek nehezen követhetők Emiatt a kínai nyelvnek olyan eltérő tájszólásai alakultak ki, hogy ma már szóban meg sem érti egymást két távoli tájról érkező kínai. E miatt nekik nehéz lenne áttérni latin írásmódra. A kínai nyelv az írásában egységes.

Más a helyzet azoknál a nyelveknél – és ezek vannak többségben – amelyeknél az írás átvett technikán alapul, tehát más nyelv írását honosították. A nyelvek hangtana eltér. Az *r* máshogy hangzik egy francia, egy német vagy egy török szájából, mégis egy jelet használnak. Ha egy nyelv különböző hangjait azonosan írunk, akkor a kiejtésben is megszűnik az eltérés. A magyar nyelv írásbeli rögzítésénél például nem használtak különböző jelet a magas és mély hangrendű *i*-re (szemben a törökkel, ahol jelölik). Kiejtésben nincs is már semmi eltérés. A nyílt *e* és az ajakkerekítéses *ë* betűnek sincs megkülönböztető jele. Ma már egyre kevesebben tudják érzékeltetni, pedig nyelvünkben ennek is van jelentősége.

1.4 A nyelvek osztályozása

Ma a világon több mint 5000 nyelvet beszélnek. Pontos számot nyilván nem lehet megadni, mert nagyon sok esetben eldöntetlen, hogy egy-egy embercsoport nyelve önálló nyelv-e, vagy csak nyelvjárás, s az is lehetséges, hogy még léteznek eddig fel nem fedezett nyelvek. Hogy ez a sok nyelv egy közös ősnyelvből származik, vagy a Föld különböző pontjain egymástól függetlenül alakult ki több ősnyelv, megválaszolatlan kérdés, de ha egy is a forrása minden földi nyelvnek, akkor is lényegesen eltérnek egymástól.

A nyelveket nem az írás mikéntje különbözteti meg. Az ukrán és a lengyel jobban hasonlít egymásra, mint a német és a lengyel. A szerb és a horvát sem tér el túlságosan, pedig az egyik cirill, a másik latin betűkkel ír. A kínai a japántól lényegesen eltér, pedig rengeteg közös írásjelük van. A nyelveket vagy kialakulásuk, rokonságuk alapján osztályozzák, vagy pedig szerkezetük szerint, ami persze összefüggésben van a nyelvek származásával.

A nyelvben található két legfontosabb szerkezet a szó- illetve a mondatszerkezet. A szavak szerkezete szerint a két alapeset, hogy nincs szerkezete (a szavak csak szótári alakjukban szerepelnek a szövegben) vagy változnak a mondatban elfoglalt szerepük szerint. Ez utóbbiakat flektáló nyelveknek nevezzük. A flektáló nyelveknél még lényeges, hogy a változás a szó belsejében megy végbe (inflexió), vagy elő- (prefix) illetve utótagokat (postfix), toldalékokat használunk Ilyen nyelvek a ragozó nyelvek. Ez utóbbi esetben még érdekes, amikor a toldalékok több, esetleg nem meghatározott számú rétegben kapcsolódhatnak a szótőhöz. Ezeket a nyelveket agglutinatív nyelveknek hívjuk. A nyelvek ilyen besorolása nem matematikai. Tehát ha egy nyelv nem nagyon ragoz, akkor nem nevezzük ragozónak. Pl. az angolban vagy a kínaiban kevés a szó változása. Pl. a 英 jelentése Anglia, a 英言 jelentése angol (nyelv), de egyik szónak sincs esetragja, mint a magyarban. Emiatt mind az angolt, mind a kínait alig lehet ragozó nyelvnek mondani. Az arab és a héber szavaiban a szó előtt és után meghatározott 1-2 toldalék lehet. Eközben maga a szótő belsejében változhatnak a magánhangzók (melyet ritkán jelölnek). Így e két nyelv két oldalról toldalékoló, egyben inflexiót használó nyelv. A török és a magyar nyelv ezzel szemben alapvetően a szó végére ragasztgatja a toldalékait több rétegben (*méreg+telen+ít+het+etlen+ül*), ezért e nyelvek agglutinatívok.

A mondat szerkezete szempontjából az alapvető, hogy a mondat részeinek a sorrendje mennyire kötött. Ennek alapján van szabad szórendű nyelv, és kötött szórendű nyelv. A kötött szórendű nyelvek között még lényeges, hogy a szerkezetileg összetartozó elemek egymás mellett helyezkednek el, vagy sem. Ha igen, akkor izoláló nyelvről beszélünk. Az izoláló nyelvek szintaxisának leírására a környezetfüggetlen nyelvtanok alkalmazása kecsegtető, hiszen a CF nyelvtanok pont ilyen szerkezetű nyelvek leírására találták ki. Az osztályozás itt sem matematikai szabatossággal értendő. Ha egy nyelvben alapvetően sorrendiség határozza meg a szavak, mondatrészek egymáshoz való viszonyát, akkor kötött szórendűnek nevezzük, míg ha a szabályok nagyobb része megengedi a mondatrészek csereberéjét, akkor szabad szórendűnek mondjuk. Ebben az értelemben lehet a magyart szabad szórendűnek, az angolt, a kínait kötött szórendűnek mondani.

A szószerkezet összetettsége és a mondatszerkezet típusa egymással szoros viszonyban van. Azok a nyelvek, amelyekben a szó nem ragozódik, a szavak mondatbeli szerepét csak a szavak sorrendjével lehet kifejezni, míg az erősen ragozó nyelvekben szabadabb a szósorrend.

2. A nyelv alapegységei

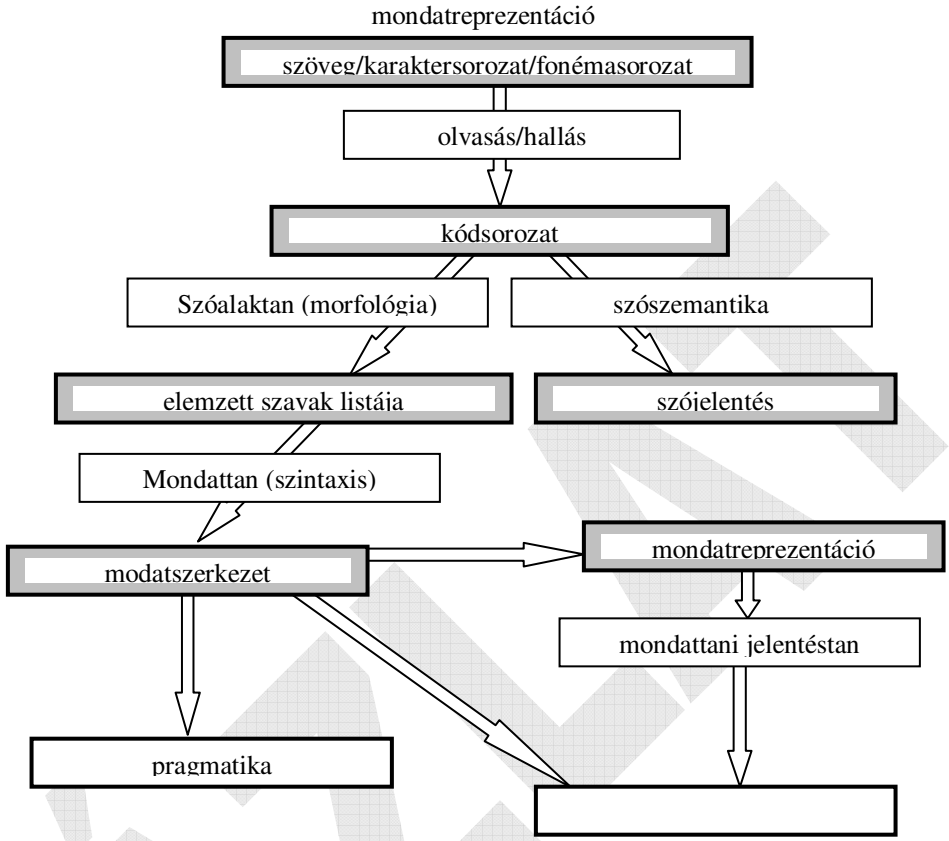
Az írott szöveg betűkből és írásjelekből áll. A beszélt nyelv ennek megfelelően hangokból, fonémákból.

Ezekből alkotott jól elválasztható egységek a szavak. A szavak esetleg több részre bonthatók. Itt nem csak a szóösszetételekre gondolok, hanem egy szó valódi funkcionális darabokra bontható. Ezek a morfémák. A magyarban morféma a szótő, a ragok, a jelek a képzők stb. Ezeknek megvan az önálló nyelvtani szerepük. Bár tudjuk, mi a szó, mégis, ha pontos meghatározást akarunk mondani, akkor bajban vagyunk. Az írott szövegben jól elkülöníthető részek. Beszélt nyelvben az egymást követő szavak összeolvadhatnak, ezért nehezebb megállapítani, a mondatnak mely része önálló szó. Az írásban is lehetnek kérdések e téren. Tudjuk, hogy a szóköz jó szeparátor, és az írásjelek általában nem a szó részei. A pont viszont lehet egy rövidítés utolsó karaktere (*magy. kir. posta*). A kötőjel is kérdéses lehet (*lót-fut*). Más nyelvekben ettől különböző jelek is okozhatnak gondot. Ha a számjeggyel írt számokat is szónak tekintjük, akkor a tizedesvessző egy szóalkotó karakter (3,141 597 8), míg a felsorolásnál használt vessző önálló egysége a mondatnak.

A mondatok szavakból és írásjelekből állnak. A szószint és a mondatszint között helyezhetjük el a szintagma fogalmát is. A magyar nyelvben szintagmára példa lehet egy névszói kifejezés (pl. névelős jelzős ragozott főnév), vagy egy ige a hozzátartozó segédigével, esetleges tagadószóval együtt. Hogy hol kezdődik, hol végződik egy mondat magyar nyelvben, abból állapítjuk meg, hogy milyen írásjeleket használunk, illetve a mondat első szavát nagybetűvel kell kezdeni. Ez persze megint nem minden esetben határozza meg a mondatok közti határt. Ha a mondat utolsó szava rövidítés, vagy… (hármas pont), akkor utána nem kell újabb pontot tenni a mondat végére. Ha persze kisbetűvel, vesszővel folytatódik a szöveg, akkor lehet tudni, hogy nincs vége a mondatnak. (*Jött, látott…, magabiztos volt, mint Július Cézár.*) Ha viszont nagybetűvel, akkor csak akkor lehetünk biztosak, ha a következő szó nem tulajdonnév. (*Jött, látott… Magabiztos volt, mint Július Cézár.*) Ha tulajdonnév a pontot követő első szó, akkor viszont annak eldöntése, hogy a kérdéses pont mondatvéget jelent-e, nehezen megválaszolható formális eszközökkel. (*Jött, látott stb. Július Cézár.*)

A mondatok felett is léteznek szintek, sőt az egyes szinteket különböző szempontok szerint lehet megközelíteni. A megközelítés lehet tisztán formai, de lehet jelentéstani.

A különböző szintek a nyelvészet különböző ágai foglalkoznak. Számunkra a következő egymásra épülő ágakat kell vizsgálni:



A szóalaktan a szavak szerkezetének, szófaji beosztásának, a ragozásuknak, a szóösszetételeknek leírásával foglakozik. A szószemantika a szavak jelentését írja le. A mondattan a mondatok szerkezetét definiálja. A mondattani jelentéstan a mondat szerkezetéből és a szószemantikából kiindulva határozza meg a mondat valamilyen szemantikai reprezentációját. A pragmatika a közlő szándékának meghatározásával foglalkozik, mely eltérhet a mondat egyszerű szemantikai jellemzésétől. A diskurzusreprezentáció a mondatok egymásutánjából von le újabb következtetéseket. További szintek is lehetségesek, de ma ezek a legfőbb elemei egy nyelvi rendszernek.

3. Betűk

3.1 Betűírás – ideografikus írás

A mai nyelvekben is használnak szóírást. A kínaiak és a japánok (kandzsi) írásjeleinek száma emiatt nagyságrendben 20 000. Ezek elsajátítása nem egyszerű feladat. Nyelv szókészletének felel meg, de nyilvánvaló, hogy ezzel a készlettel kell leírni az új fogalmakat is. A jövevényszavakat, új fogalmakat a meglévőkből rakják össze annak alapján, hogy kiejtése hasonlítson az összerakott szavak egymásutánjához. A japán erre a célra más karaktereket használ, a katakana és hiragana jeleket, amely a szótagírásra alkalmas. A szótagíráshoz párszáz karakter elegendő. A betűírás a leggyakoribb. Ebben az esetben egy nyelven 20-100 karakterrel mindent meg lehet adni.

Az írás iránya nyelvenként eltér. A hagyományos kínai írásban a jeleket fentről lefelé (sorok), jobbról balra (oszlopok) írták. Ma már hozzánk hasonlóan balról jobbra, és a sorokat fentről lefelé írják. A betűíró népeknél egyesek jobbról balra, mások balról jobbra vetik a betűket. Ennek praktikus okai lehettek. A papírlapra írók – feltételezve, hogy jobbkezesek – azért írnak balról jobbra, hogy írás közben lehessen látni a már kész szöveget. Ha viszont tekercsre rótták a sorokat, célszerűebb volt a jobbról balra írás, mert így jobb kézzel könnyebb volt közben továbbtekerni a tekercset. Ilyen esetben lehet ideális a fentről lefelé írás is, amit a tibetiek ma is használnak. A lényeg csupán az, hogy a jelek sorrendje egyértelmű legyen. Az ókori görögök

egy időben váltott irányban írtak: egy sort jobbra, egy sort balra, mint az ökörhugyozás. Így az olvasás tűnik könnyebbnek. Nem kell keresni, hol a következő sor eleje.

Példa ideografikus (kínai) írás jeleire: 世 界 挪 威 葡 萄 牙 保 加 利 亚

Példa szótagírás (japán katakana) jeleire: エスペラント ポルトガル チェコ スロバキア人

3.2 Karakterek, betűk írásjelek, digráfok, trigráfok…

Bár a betűírás eredendően a fonetika írott képe, ma már nem egyértelmű, hogy milyen hangot mivel jelöljünk. Nem csak arról van szó, hogy a *j* hangot írhatjuk *ly*-nal és *j*-vel is (persze a szó meghatározza, mit kell használni), hanem azt is, hogy egy leírt szó betűi kiejtéskor változhatnak. A *vasgyúró* kiejtve *vazsgyúró*, a *teljes*t meg *tejjes*nek ejtjük. Angol szövegben már nem is lehet követni pontosan, mit hogyan ejtünk. A *put* ige kimondva is *put*, de a *but* szó kiejtve *bât*.

A legtöbb európai népnél az írásbeliség elterjedése a reformációhoz (a biblia fordításához) kötődik. Korábbról is maradtak fenn írásos emlékek, de azok ritkák voltak. A korai katolicizmus idején a hivatalos írott nyelv a latin volt, míg más – a nemzethez kötött – írásmódok később kivesztek. A latin betűkkel nem lehetett minden hangot ábrázolni, ezért azt bővíteni kellett. Ennek két módja volt. Az egyik, hogy a latin karaktereket kiegészítették módosító jelekkel, ékezetekkel. A másik lehetőség, hogy több karakterrel jelöltek egy betűt. A csehek a korábbi módszert választották. Mi magyarok, illetve az írást nálunk bevezető egyházfiak – később nyomdászok – a magánhangzók bővítésére az ékezeteket (diakritikumokat) használták, míg a mássalhangzóknál kettős, hármas jelekkel, digráfokkal, trigráfokkal oldották meg a bővítést:

a á b c cs d dz dzs e é f g gy h i í j k l ly m n ny o ó ö ő p q r s sz t ty u ú ü ű v w x y z zs

és ezek nagybetűs változatai. Ebben a listában szereplő q, x és y önálló betűként ritkán fordul elő. Viszont még más, hivatalosan nem magyar betű is előfordulhat magyar szövegben: *ä jäger*, *ch technika*, *th Batthány*… A mássalhangzóknak létezik nyomatékos (hosszú) párja is. Ilyenkor kettős, hármas betűknek csak az első karakterét kell duplázni. Tehát a betű és a karakter két külön fogalom. A betű egy vagy több karakterből áll. Néha nehéz megfogalmazni, hol a határ. A sémi nyelvekben, mint említettem, a magánhangzók jelölése opcionális. Tulajdonképpen, ha jelölik, akkor kiegészítője az őt megelőző mássalhangzónak. Írásban úgy jelenik meg, mint magyarban az ékezet egyes magánhangzókon. A mássalhangzó ezen kívül kiegészíthető módosító jellel, arabban hangsúllyal is. Hogy ezek a jelek a mássalhangzó részét jelentik, vagy önálló karakterek, az már hozzáállás kérdése. A héber német jelentésű (*germanit*) szó mássalhangzói alatt levő pöttyök jelentik a magánhangzókat:

גֶּרמָנִית

A betűről már tudjuk, hogy az adott nyelv ábécéjében felsorolt dolgok, melyek karakterekből állnak. De mi a karakter? A karakter alapvető egysége az írásnak. A nyomdász ezekből rakja (szedi) a szöveget, mikor írunk, ezek az alapvető egységek, amiket egymás után a papírra vetünk, az írógép billentyűjén leütünk. Tulajdonképpen formája van, bár eltérhet. Mégis ugyan azt a betűt jelentik a következők:

ɑ a a a **a** *a* a *a* *a* ***a***

A forma absztrakciója jelenti a karaktert. Ezzel szemben, ha egy bolgár vagy orosz szövegben jelenik meg ugyanez a forma, akkor más karakternek számít, mint a magyarban, de a nagy *o* betű sem tér el különösen a nullától. (*O*, *0*)

Magyarban minden betűnek létezik kis- és nagy változata. Az arab nyelvben viszont egy betűnek két, három, esetleg négy alakja is lehet, aszerint hogy a szó elején, közepén vagy végén helyezkedik el, netalán egyedül áll. Példánkban ugyanaz a betű jelenik meg (jobbról balra nézve a szavakat) önállóan, a szó végén (baloldalán), a szó közepén (második karakterként), és a szó elején (jobboldalán).

ي بجي يبج

A magyarban, ha csak a kisbetűket számolom, 44 betű van, melyet 35 különböző karakterrel le tudok írni. A szövegben nem csak betűk, hanem számjegyek és írásjelek is találhatók. A magyar írógépen a következő jeleket lehet még használni: *' „ " + ! % / = ( ) , ? . : - _* és a szóköz. Érdekesség hogy írógépen nincs *í* (hosszú *i*), *0* (nulla) és *1* (egyes). Ezeket *i'* jelekből rakják össze, *O* (nagy *o*) és *l* (kis *L*) karakterekkel helyettesítik. Mint karakterek (formák) meg is felelnek a célnak, de a számítógépen tarthatatlan, hogy különböző jeleknek azonos legyen a belső ábrázolása (a kódja). Számítógépen még számos, ezektől eltérő karakter ábrázolható, melyről szabályokat is alkottak a nyelvészek. Így megkülönböztetendő a rövid és a hosszú kötőjel, létezik « és » jel, ha idézőjeles mondatban akarunk idézetet használni, stb.

Ha a kis- és nagybetűket is különbözőnek tekintjük, akkor sem haladja meg a 200-at a használt jelek száma. Ez így van minden betűírásos nyelv esetén. Szótagírásnál a nyelv karaktereinek száma pár százra tehető, míg szóírásnál a több ezer karakter a nyelvre jellemző.

3.3 Kódolások

A kódolás a digitális technika következménye. A karakterek formája ilyenkor lényegtelen. A fontos, hogy egyértelmű számot rendeljünk minden karakterhez. Hogyan kódoljuk számítógépen az írást? Természetesen úgy, hogy ha egy másik gépen használom a szöveget, akkor ott is ugyanazt a karaktert jelentse. A számítógépes szövegábrázolás a távadatátvitelből nőtt ki. Ezért kezdetben (ötbites) teletex kódokat feletettek meg a karaktereknek. Ez persze arra is alig volt elegendő, hogy a nagybetűs angol ábécé szövegeit kódolják. Emiatt már a számjegyek írásához úgynevezett váltókódot kellett alkalmazni. Amikor a szövegben számjegy vagy különleges karakter szerepelt, előtte egy váltókódot kellett küldeni, majd a számjegyes rész után egy betűváltóval vissza kellett állítani, hogy újra szöveget lehessen kódolni. Ezen a kissé nehézkes módszeren csak a több bites ábrázolás segített. Egyrészt a lyukkártyás rendszerek EBCDIC (Extended Binary-Coded Decimal Interchange Code) kódja, másrészt az Amerikában gyorsan elterjedő ASCII (American Standard Codes for Information Interchange) vetélkedett egy ideig egymással. A kis- és nagybetűk már nem okoztak gondot, de kizárólag ékezet nélküli szöveget lehetett kódolni. Az ASCII pl. hétbites ábrázolást nyújtott. A nyolcadik bitet keresztparitásra, hibaészlelésre használták. Az alacsony kódokat vezérlőjeleknek használták, így 2^7-33=95 különböző látható jelet lehetett ábrázolni. Hosszú távon az ASCII győzött. Ma a legtöbb számítógépes kódolásnak ez az alapja. Mivel a számítógépek legnagyobb gyártói és felhasználói amerikaiak voltak, megelégedtek az angol szövegek ábrázolásával.

A gépek tömeges felhasználását a PC-k megjelenése tette lehetővé a 80-as évek elején. A nemzeti sajátos betűk használata ekkor elengedhetetlenné vált. Persze az Egyesült Államokban is gyakoriak a más nyelvű szövegek. Északi szomszédjaik franciául, délre pedig spanyol nyelven beszélők laknak. Az európai országokkal is fontos volt a kapcsolat. Ennek érdekébe kiterjesztették a hétbites kódokat nyolcbitesre. Az így kiterjesztett kódtáblázat (melyet később 437-esnek neveztek) tartalmazta az összes nyugat-európai nyelv betűjét. Kis számolással kiderül, hogy nyolc biten reménytelen az összes betűírásos karaktert kódolni. A latin betűkön kívül léteznek cirill, héber, görög, arab és még számos más, ezektől lényegesen különböző írásmód. Ezeknek mind eltérő a kiterjesztésük egymástól. Sajnos a latin (jellegű) írásnak sem fér el az összes betűje 256 elemből álló kódlapra. Nem csak magánhangzóknak létezik diakritikuma (különböző ékezetes variánsa), hanem a mássalhangzóknak is lehetnek kiegészítő jelei:

aàáâãäāăąạảấầẩẫậắằẳẵặ ccĉċčç ţţťŧ sśŝşš …

Ha mindet felsorolnánk, meglepően nagyméretűnek adódna a nemzetközi ábécé. Ezért aztán nyelvi csoportokat érdemes készíteni. Pontosabban kódlapokat. Egy kódlapra több nyelv karakterei is rákerülhetnek.

Különböző cégek eltérően bővítették ki a hétbites ASCII-t. Az így kialakult nyolcbites kódlapokat utólag kódlapszámmal azonosították. Néhány példa:

A Microsoft DOS-os kelet-európai kódlapja, a 852-es

	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0
0	szóköz	0	@	P	`	p	Ç	É	á	•	•	•	Ó	feltételes elválasztójel
1	!	1	A	Q	a	q	ü	•	í	•	•	•	ß	˝
2	„	2	B	R	b	r	é	•	ó	•	•	•	Ô	˛
3	#	3	C	S	c	s	â	ô	ú	•	•	Ë	•	ˇ
4	$	4	D	T	d	t	ä	ö	•	•	•	•	•	˘
5	%	5	E	U	e	u	•	•	•	Á	•	•	•	§
6	&	6	F	V	f	v	•	•	Ž	Â	•	Í	Š	÷
7	’	7	G	W	g	w	ç	•	ž	•	•	Î	š	¸
8	(	8	H	X	h	x	ł	•	•	•	•	•	•	°
9	)	9	I	Y	i	y	ë	Ö	•	•	•	•	Ú	¨
A	*	:	J	Z	j	z	•	Ü	¬	•	•	•	•	˙
B	+	;	K	[	k	{	•	•	•	•	•	•	•	•
C	,	<	L	\	l	|	î	•	•	•	•	•	Ý	•
D	–	=	M	]	m	}	•	Ł	•	•	•	•	Ý	•
E	.	>	N	^	n	~	Ä	×	«	•	•	•	•	•
F	/	?	O	_	o	•	•	•	»	•	¤	•	´	nem törő szóköz

A Microsoft Windowsos kelet-európai kódlapja, az 1250-es

	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0
0	szóköz	0	@	P	`	p	•	nincs	nem törő szóköz	°	•	•	•	•
1	!	1	A	Q	a	q	nincs	‘	ˇ	±	Á	•	á	•
2	„	2	B	R	b	r	,	’	˘	˛	Â	•	â	•
3	#	3	C	S	c	s	nincs	“	Ł	ł	•	Ó	•	ó
4	$	4	D	T	d	t	„	”	¤	´	Ä	Ô	ä	ô
5	%	5	E	U	e	u	...	•	•	µ	•	•	•	•
6	&	6	F	V	f	v	†	–	¦	¶	•	Ö	•	ö
7	’	7	G	W	g	w	‡	—	§	·	Ç	×	ç	÷
8	(	8	H	X	h	x	nincs	nincs	¨	¸	•	•	•	•
9	)	9	I	Y	i	y	‰	™	©	•	É	•	é	•
A	*	:	J	Z	j	z	Š	š	•	•	•	Ú	•	ú
B	+	;	K	[	k	{	‹	›	«	»	Ë	•	ë	•
C	,	<	L	\	l	|	•	•	¬	•	•	Ü	•	ü
D	–	=	M	]	m	}	•	•	feltételes elválasztójel	˝	Í	Ý	í	ý
E	.	>	N	^	n	~	Ž	ž	®	•	Î	•	î	•
F	/	?	O	_	o	nincs	•	•	•	•	•	ß	•	˙

A Macintoshon használt kelet-európai kódlap, a 10029-es

	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0
0	szóköz	0	@	P	`	p	Ä	•	†	•	•	–	•	•
1	!	1	A	Q	a	q	•	•	°	•	•	—	Š	Û
2	„	2	B	R	b	r	•	í	•	•	¬	“	,	Ú
3	#	3	C	S	c	s	É	•	£	•	✔	”	„	•
4	$	4	D	T	d	t	•	•	§	•	•	‘	š	•
5	%	5	E	U	e	u	Ö	•	•	•	•	’	•	•
6	&	6	F	V	f	v	Ü	•	¶	•	•	÷	•	•
7	’	7	G	W	g	w	á	ó	ß	•	«	•	Á	•
8	(	8	H	X	h	x	•	•	®	ł	»	•	•	ý
9	)	9	I	Y	i	y	•	ô	©	•	...	•	•	Ý
A	*	:	J	Z	j	z	ä	ö	™	•	nem törő szóköz	•	Í	•
B	+	;	K	[	k	{	•	õ	•	•	•	•	Ž	•
C	,	<	L	\	l	|	•	ú	•	•	•	‹	ž	Ł
D	–	=	M	]	m	}	•	•	¨	•	Õ	›	•	•
E	.	>	N	^	n	~	é	•	•	•	•	•	Ó	•
F	/	?	O	_	o	nincs	•	ü	•	•	•	•	Ô	ˇ

A fenti megoldásokban közös, hogy a kisebb kódú betűk megegyeznek a hétbites ASCII kódokkal, valamint mindegyik tartalmazza a teljes magyar karakterkészletet. Az egyéb karakterek már lényegesen eltérnek.

A 852-es és a 1250-es román lengyel, cseh és szlovák karaktereket is tartalmazza, a 100029-es pedig a balti népek különleges ékezeteit is fedi. Egyikben sincs cirill betű, de a törököknek sem elegendő a készlet.

Még számos olyan kódlap létezik, amely alkalmas az összes magyar karakter ábrázolására. Amikor még a nyugati cégek nem gondoltak ránk, különböző műhelyekben különbözőképpen oldották meg a magyar írást. Csupán két olyan betűnk van, amely más nyelvekben nem lelhető meg. Ez a hosszú *ő/Ő* és a hosszú *ű/Ű*. A korai megoldások közül had említsem meg a Wordperfectben, a Ventura kiadványszerkesztőben használt kódkiosztást. Országosan a legelterjedtebb az úgynevezett CWI-2-es kódlap volt, amely a PC-kbe beégetett 437-es kódlaptól alig tér el.

A hét bit szűkösségéből más úton is ki lehet lépni Az egyik mód a repülő ékezetes írás. Ekkor a kérdéses karaktert megelőzi (vagy követi) egy módosító karakter. Ha nem okoz félreértést, akkor a módosító karaktert együtt lehet értelmezni az őt megelőző (követő) betűvel:

Az u'tto:ro" segi't le'pten, nyomon = *Az* úttörő *segít lépten, nyomon*

Ennek egy érdekes változata a Prószéky kód. Ha egy betűt szám követ (feltételezve, hogy nincs olyan szöveg, amelyben a betűk és a számok vegyesen lennének egy magyar szóban), akkor a karakter megfelelő sorszámú diakritikumát kell érteni. A sima ékezet az 1, az umlaut a 2 és a hosszú kettős ékezet a 3. A fenti mondat a következőképpen néz ki Prószéky kódban:

Az u1tto2ro3 segi1t le1pten, nyomon

Ezt a kódot ma is használják a Magyar Tudományos Akadémia Nyelvtudományi Intézetében, de a módosító számok nem állnak meg 3-nál. Régebbi irodalmunk gépbevitelénél jelölni kellett azt is, hogy a több száz évvel ezelőtti írásban (nyomdában) hogyan jelölték a betűket. Mivel akkor még nem volt egységes az írásunk, ezek eltértek egymástól.

A repülő ékezethez hasonló kódfüggetlen megoldást választott Donald E. Knuth is, a TEX megalkotója. A különleges karaktereket, illetve azokat a kiegészítőket, melyek módosítják a betűket, kulcsszavakkal jelölik, és precíz formalizmussal illesztik a szövegbe. TEX-ben a fenti mondat így néz(het) ki:

Az \'utt\"or\h o seg\`\i t l\`epten, nyomon

A kódfüggetlen írásmód nagy előnye, hogy Unixban, DOS-ban, Windowson azonos a szöveg értelmezése. Hátránya, hogy nehezen olvasható, illetve megjelenítő, konvertáló program kell a természetes olvasási kép megkapásához.

Megjegyzem, hogy lehetséges olyan kódolás, amely elszakad az írástól, viszont nyelvészi szempontból hasznos lehet. Pl. magyarban lehet(ne) kódolni a szövegben, hogy magas vagy mély hangrendű az *i* betűről (pl. *i*, *ï*) vagy nyílt illetve zárt *e* betűről van szó (*e*, *ë*). Megemlíteném, hogy a 90-es években készült Lektor (Seregy Lajos, Vanczák János, Hámori Miklós, Béres Tamás) helyesírás-ellenőrző belső ábrázolásában nem létezik digráf, trigráf. Ezekre, az ő „humán kódolásukban" külön kódot vezettek be.

Kódlapok alkalmazásánál a vegyes nyelvű szövegekben nem biztos, hogy mindegyik nyelv betűi elférnek egy nyolcbites kódolással. Ebben az esetben a szöveg tárolt részében el kell helyezni kódlapváltó karaktereket. Ezt a megoldást választotta a ma már elfeledett ChiWriter, a WordPerfect korábbi változatai, de a 16 bites Microsoft termékek (Word 6) is így tudták kódolni a szövegeket, ha francia, magyar, görög vagy orosz nyelvű szavak vegyesen voltak egy dokumentumon belül. Mivel egy szövegen belül ritkán változtatják a nyelvet, ez nem jelent jelentős memóriatöbbletet. Ma már olcsóbb a memória. A jövő minden bizonnyal a UNICODE-é, vagyis a szabványosított karakterkódolásé. Nagy előnye, hogy ebbe a rendszerbe a Föld minden betűjének egyedi helye van. Sőt erre az univerzális kódlapra már nem csak a betűírásos nyelvek férnek rá, hanem a szótagírás és az ideografikus írás jelei is kényelmesen elférnek, hisz több mint 100 000 különböző jelet lehet kódolni. Ha a szöveg alapvetően latin írású, akkor gazdaságos az UTF-7-es, UTF-8-as kódolás. Ha a karakter UNICODE szerinti kódja elfér 7 biten, akkor azt egy-egy bájtban tárolják, s a legfelső bit természetesen 0. Azon ritkább esetben, ha nagyobb a kód, Akkor a hosszabban ábrázolt kód bájtjaiban binárisan a legfelső bit mindenütt 1-es. Annyi bájtból áll a kód, ahány egymás utáni bit egyes az első bájtban. A további kódrészletekben is 1-es a legfelső bit, és a második pedig 0. Így ebben az esetben 6 információs bitet tartalmaznak a UNICODE értékéből a további bájtok.

	bináris 32 bites ábrázolás	UTF-8
$<2^7$	00000000 00000000 00000000 0xxxxxxx	0xxxxxxx
$<2^{12}$	00000000 00000000 00000xxx xxxxxxxx	110xxxxx 10xxxxxx
$<2^{17}$	00000000 00000000 xxxxxxxx xxxxxxxx	1110xxxx 10xxxxxx 10xxxxxx
$<2^{22}$	00000000 000xxxxx xxxxxxxx xxxxxxxx	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
$<2^{27}$	000000xx xxxxxxxx xxxxxxxx xxxxxxxx	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

Ez két bájt használata esetén 11 bit, 3 bájt esetén 16 bit, 4 bájt esetén 21 bit, 5 bájt esetén 26 bit. Ha nincs szükség több mint 2^{27} különböző értékre, márpedig a földi írásokra ez biztos elegendő, akkor ennél hosszabb kód nem kell. De ha mégis, akkor lehet tovább folytatni 7 bájtig, ami 2^{37} különböző értéket jelölhet.

A 16 bites kódkészlet nem bizonyult elégnek a földi nyelvek kódolására. Az ideografikus írásmód esetén nyelvenként (pontosabban írásmódonként) legalább 20 000 különböző jelre van szükség. Ha három, vagy annál több különböző ideografikus írásmód létezik – márpedig létezik, főként, ha a holt nyelveket is számítom – akkor kevés a 16 bit. A UNICODE készlete ma már meghaladja a 65365-öt. Ezért az UTF-16-os kódolásnál (16 bites alapkódot használva) is lehetőség van nagyobb értékű kódok többszavas ábrázolására, de ez eltér az UTF-8-as kódolás logikájától.

Ha az információ tömörsége lényeges, akkor nem bájt vagy szó az alapegység, hanem bit. Változó hosszúságú kódokat használnak a Huffman-kódolásnál, ami biztosíthatja statisztikailag legrövidebb tárolását a szövegnek. Ennél tömörebb kódolás csak akkor lehetséges, ha nem karakterenként, hanem esetleg karaktercsoportonként kódolunk. (pl. zip. rar…)

3.4 Kódkonverziók

Az egyetemi években rossz jegyzetelő voltam. Emiatt rászorultam diáktársaim írásainak használatára. Az egyik társam, Nagy Zsigmond remekül jegyzetelt, de betűit rajta kívül csak én tudtam dekódolni. Ez volt a szerencsém.

A különböző kódolások ennek ellenére még sokáig fennmaradnak. Ennek nem csak történeti okai vannak, hanem egyes alkalmazásoknál számít, hogy az információ mekkora területet foglal el. Gondoljunk a miniatürizált szótárakra, vagy arra, hogy lassabb adatátvitelnél nem mindegy, hogy mennyi ideig tart az információtovábbítás. Ha erre nem is gondolunk, még évtizedekig lesz archívumunkban korábban írt anyag, mely nem az egységes UNICODE-ban készült. Emiatt szükség lehet különböző konverterekre. A szövegfeldolgozás folyamán célszerű, ha rögzített hosszúságúak a kódok, de a tárolásnál nem. Egy kódolási rendszer akkor jó, ha könnyedén lehet átírni az egyikből a másikba a szöveget. Nem feltétlen szükséges, hogy azonos hosszúságú legyen minden karakter kódja, de ajánlatos az úgynevezett prefix kódolás. A másik követelmény, hogy mindkét irányba egyértelmű legyen a kódolás. Ez általában nem tartható. Szinte minden kódrendszer tartalmaz olyan karaktereket, amit a másik nem. Ilyenkor a minimális követelmény, hogy a kérdéses nyelv betűit egyértelműen lehessen az egyik kódrendszerből a másikba vinni. Ilyen értelemben könnyen lehet konvertálni 852-es és 1250-es magyar szöveget egyikről a másikra, pedig mindkettőben létezik olyan karakter, amelyik a másikból hiányzik. Ha a karakterek forráskódjának hossza állandó, az átkódolás egyszerű táblázattal (tömbbel) leírható. Algoritmikusan mindenképpen ez az ajánlott, hisz a számítógép egyik leggyorsabb művelete, ha egy tömbből index alapján kiemelünk egy értéket. A fenti megszorításokkal a visszakódolás is táblázat (tömb) segítségével elvégezhető. A „távirati stílus” kódjánál csak néhány ékezetes karaktert kell átírni: á = aa, é = ee, í = ii, ó = oo, ö = oe, ő = ooe, ú = uu, ü = ue, ű = uue

A szöveg kódolása egyértelmű, de visszakódolása már nem feltétlen. Gondoljunk a *koordinátor* vagy a *faajtó* szóra.

A kódolásnak egyszerűnek és gyorsnak kell lennie. Gyakorlati követelmény, hogy a kódolást-dekódolást véges állapotú fordítóval el lehessen végezni. A táblázatos megoldás is ilyen, ami mindig járható állandó hosszúságú kódoknál, de pl. UNICODE-ról 8 bites kódra való konverziónál nem gazdaságos 65 kilobájt fenntartása. A fent említett kódlapos megoldások ezt az egyszerű követelményt teljesítik. A repülő ékezetes kódolás, a TEX kódok kódlapos átírása is mindig kódolható véges állapotú fordítóval, de nem feltétlenül determinisztikussal. Az említett esetekben viszont egy előretekintéssel – ami a véges állapotú automaták elméletében nem szokványos – determinisztikussá tehető a konvertálás.

Az egyszerű táblázatos kódolás feltételei nem, mindig állnak fenn. Gyors a kódolás akkor is, ha „regiszterváltós” a kódolás, mint pl. a teletextnél. A UNICODE-nál nincs elvi határa a karakterek kódjának. Korábban abban a hitben teremtették meg, hogy a világon nincs szükség 65 000 különböző írásjelre. A gond nem is

a betű, illetve a szótagírásnál jelentkezett, hanem az ideografikus jelek kódolásánál. Ha feltételezzük, hogy egy ideografikus írásmódhoz nagyságrendben 20 000 különböző jel kell, akkor három különböző készlet már 60 000 kódot elhasznál a lehetségesek közül, ezért nincs remény arra, hogy elegendő legyen a világ összes írásmódját 16 bites rögzített hosszúságú kóddal reprezentálni. Emiatt a UNICODE reprezentáció nem célszerű állandó hosszúságú kódokkal megoldani. Erre ad megoldást az UTF-8 illetve az UTF-16-os kódolás. A nem állandó hosszúságú kódolásnál célszerű a prefix jellegű kódolás. Prefix a kódolás, ha egy egységet (jelen esetben karaktert) egyértelműen azonosíthatunk anélkül, hogy akár egy elemmel továbbolvasnánk a sztringet. Más szavakkal, egy egység sem lehet valódi kezdete egy másiknak. Ha véges állapotú (fordító) automatával kódolunk, akkor prefix kódolás esetén megtehetjük azt egy determinisztikus véges fordítóval, így a bemenettel szinkronban jelenik meg a megfelelő kód a kimeneten.

3.5 Egyértelműség

Ha jól tudom, a magyar rovásírás betűi még nem kaptak helyet a UNICODE tárházában. Nem, mert a két nagy tábor – melyik az igazi ékírás – nem tudott megegyezni.

Bár a nyelvelméletek és leírások feltételezik a karakterek, kódolások egyértelműségét, a gyakorlat néha megcáfolja ezt. A betűket, karaktereket jól definiált zárt halmazoknak tekinthetjük. Az átkódolásnál azért történhetnek meglepetések. Látszatra azonos karaktereknek lehet különböző kódjai. Pl. a cirill *а* és a latin *a* semmiben sem különbözik, de kötelesek vagyunk más kódot használni. Hasonló a helyzet a cirill *Г* és a görög *Γ* estén is. A román *ţ* betűnek kétféle változata van. Vigyázni kell, mert a kisbetű-nagybetű megfeleltetés függhet a nyelvtől: *I-i* általában, de a törököknél *İ-i*, *I-ı*. Bár a hollandoknak létezik *ij* karakterük (*Dijstra, van Dijk, Nijmegen*), helyette mégis az *i* és *j* két betűt használják. Az írógépi hagyományok miatt más egykarakteres digráfok (*œ, æ*) helyettesítődnek a két karakter egymásutánjával. A nem törő szóköz szemre megkülönböztethetetlen a szokásos szóköztől. Szóval legyünk mindig óvatosak. A karakterek és kódjai definíciója sem teljesen egyértelmű minden esetben. A nyelvészeten belül a legkevesebb többértelműség azonban a betűk terén jelentkeznek. Ha egységesen és egyszerűen akarunk kezelni korpuszokat – szöveges állományok gyűjteményét – ajánlatos az állományokat normalizálni, melynek legfontosabb része, hogy a karaktereket, betűket egyféle, egyértelmű kódra kell konvertálni.

Meglepetéssel szolgálhat, ha kihalt nyelveket akarunk megfejteni a megmaradt írások alapján. Ennek érdekes példája a maja írás. A kutatók vagy 800 különböző jelet ismertek fel, ezért abban a hitben voltak, hogy valószínűleg szóírás hieroglifáira leltek. Később derült ki a turpisság. A nép játékos és művészi hajlamú írástudói egy karakternek számos kaligráfját használták még egy íráson belül is. Szótagírás volt, de egy karakternek a fő tulajdonságait megtartva sok alakja létezhetett.

Feladat 1: Nézze meg, a különböző változó hosszú kódolások melyike prefix jellegű, illetve melyeket lehet LM (a hosszabbik találat) szerint értelmezni.

Feladat 2: Készítsen táblázatos oda-vissza kódkonvertálót, rugalmas, könnyen megadható kódtáblával.

Feladat 3: Készítsen konvertálót változó hosszú kódokat is kezelő esetre is.

3.6 Magánhangzók, mássalhangzók

Mint említettem, a magyar írásmód követi a kiejtést, tehát a leírt betűk lényegében a fonetika tükrét jelentik. A betűk két osztálya a magán- és a mássalhangzók. Nevük alapján az önállóan kiejthető hangok írásos képei a magánhangzók, melyeket pedig magánhangzó társaságában mondunk ki, azok a mássalhangzók. Persze ez igazán megállapodás kérdése. Sok olyan betű van, melyet mássalhangzónak tartunk, mégis gond nélkül ki tudunk mondani. Ilyenek az *s, sz, l, r, m*, de a *t, d, k* már nem jön ki szánkon önmagában. Hogy mennyire konvenció kérdése, azt a cseh vagy a szlovák nyelv mutatja. A fagylaltot jelentő *zmrzlina* szó szerintünk kiejthetetlen, szerintük tele van magánhangzóval. Egy hottentottának a kattogó, pattogó hangok könnyedén kiejthetők önmagukban is.

Mind a magánhangzókat, mind a mássalhangzókat lehet tovább osztályozni. A mássalhangzók legfontosabb tulajdonsága, hogy az *r*-et, a *j*-t, az *l*-et, az *m*-et, az *n*-et, az *ny*-et, az *ly*-t és a *h*-t kivéve mindegyiknek van zöngés, illetve zöngétlen párja. Soknak van kemény, illetve lágy párja. Ezenkívül osztályozzák a kiejtés hossza, és a képzés helye az ajkak állása stb. szerint.

A magánhangzóknak is megvan a tulajdonságuk. Alapvetően mindegyiknek van rövid és hosszú változata, magas és mély hangrendű párja, lehetnek ajakkerekítésesek és ajakkerekítés nélküliek, és a képzés helye szerint is lehet csoportosítani.

magas	mély
ü, ű	*u, ú*
ö, ő	*o, ó*
e, é	*a, á*
i, í	

A szépnek tűnő rendszerben a legszembeötlőbb, hogy az i, í párról nem adtam meg, hogy magas, vagy mély a hangrendje. Nem, mert írásban (és ma már kiejtésben) nem látszik semmi különbség. Ennek ellenére van. Mint később láthatjuk, a szó hangrendje alapvetően a benne levő magánhangzók hangrendjétől függ. Ennek megfelelő toldalékkal lehet csak ellátni a szót. A *szív* szavunk tulajdonképpen két szót jelenthet. Az egyik, az ige, melyben az *í* mély hangrendű, ezért ragozásában mély hangrendű toldalékot kaphat: *szívok, szívnak*… Ezzel szemben a főnévben magas hangrendű az *í*: *szívek, szívnek*… A másik érdekesség, hogy az *a*-nak és az *e*-nek a hosszú párját máshogy ejtjük ki – nem csak időtartamban – mint a rövidet. Ez persze csak tájszólás kérdése. A tarjániak kiejtésében csak megnő az *a* időtartama. Így ők az mondják, hogy *Salgootarjaan*. A magánhangzó duplázásával jelöltem a hossz növekedését.

Feladat 1: Próbálja a magyar mássalhangzókat is kevés tulajdonság alapján osztályozni.

Feladat 2: Gyűjtsön olyan szavakat, melyben egyes magyar betűkhöz a mai írásmódtól eltérő módot használnak – főként történelmi nevekben találhatók.

Feladat 3: Keressen olyan eseteket, amikor a kiejtés nem pontosan az írást követi.

3.7 Betűstatisztika – nyelvfelismerés. Markov-láncok, szöveggenerálás

A nyelvre nem csak szavai, mondatszerkezete jellemző, hanem betűkészlete is. Egyes nyelvek olyan karakterrel rendelkeznek, melyek más nyelvekből hiányoznak még akkor is, ha alapvetően azonos írásmódot használnak. Pl. csak a magyar nyelvben használnak hosszú *ő*/*Ő* és *ű*/*Ű* betűt, csak a törökben létezik pont nélküli *ı* – egyben pontos nagy *İ*, de az eszperantóban számos olyan ékezetes betű van, amely más nyelvben nem találunk meg. Ezek előfordulásánál a szöveg nyelvét könnyen azonosíthatjuk.

Ha ez nem elég, akkor ennél többet mond a betűk eloszlása a nyelvben. Közismert, hogy a magyarban a magánhangzók közül a leggyakoribb az *e*, utána az *a* betű következik, míg az *ü* betű elég ritka. Bár emberenként változó. Míg Petőfi Sándor kedvenc magánhangzója az *e* volt (*Mely nyelv merne versenyezni véled*), Arany János gonddal igyekezett az átlagnál több mélyhangrendű szót használni. Nézzük két magyar műnek a kartergyakoriságát:

János Vitéz		Toldi Estéje	
karakter	gy.	karakter	gy.
szóköz	7072	szóköz	9391
e	4622	*e*	5501
t	3986	*a*	5005
a	3791	*t*	3909
n	2708	*n*	3399
l	2648	*l*	3376
s	2520	*s*	2974
k	2150	*k*	2385
o	1980	*r*	2342
sorvég	1941	*o*	2250
á	1873	*i*	2111
r	1863	*z*	2178
z	1706	sorvég	2093
g	1661	*g*	2085
m	1645	*m*	2002
i	1606	*á*	1990
é	1312	*,*	1686
y	1188	*é*	1650
,	1054	*y*	1475
d	1028	*d*	1337
v	879	*v*	1192
b	727	*h*	990
h	695	*b*	988
j	651	*j*	883
.	572	*ö*	605
ö	522	*.*	555
u	456	*f*	524
p	406	*u*	477

f	376	*c*	470
c	362	*p*	424
ó	338	*ő*	421
ő	278	*ó*	409
"	264	*-*	339
A	203	*M*	286
M	188	*:*	257
ü	178	*"*	240
;	171	*ü*	225
J	168	*;*	206
H	161	*A*	182
í	150	*S*	172
S	137	*í*	167
ú	100	*H*	159
E	97	*N*	139
:	94	*T*	138
N	86	*E*	125
!	84	*B*	123
K	84	*!*	121
I	76	*ú*	116
?	71	*K*	108
ű	66	*?*	93
D	58	*ű*	81
-	55	*L*	80
É	49	*D*	78
T	49	*V*	76
V	43	*'*	66
L	39	*É*	66
F	36	*I*	65
B	36	*F*	61

C	25	C	43
O	21	R	40
R	21	J	36
'	19	O	32
U	17	P	23
G	11	G	19
Ő	10	U	15
P	10	Ő	15
Ő	9	Á	14

Á	8	Í	11
Í	6	(	10
(	4	)	10
)	4	*	6
4	5	Ő	5
Ó	3	Z	4
Z	2	Ü	4
Ú	2	Ú	3
Ü	1	Ó	1

A karakterek sorrendje a két oszlopban eltér, de pl. az első 8, az első 20 és az első 32 karakter megegyezik mindkét írónál. A lista vége viszont lényegesen eltér egymástól.

Ennek ellenére, a nyelvek betűstatisztikájuk alapján jól szeparálhatók. Az első betűgyakoriság-számítást a milánói Sicco Simonetta végezte 1380-ban, nem nyelvészeti céllal, hanem a titkosírás optimalizálása céljából. Morse is a betűk gyakorisága alapján dolgozta ki ábécéjét. Az angol nyelv ilyen jellegű vizsgálatát a 2001-ben elhunyt Claude Shannon végezte el 1951-ben.

Vegyünk egy-egy nagy szöveget különböző nyelveken, készítsük el a karakterstatisztikát, és az így kapott eloszlást vessük össze bármilyen bejövő levél karaktereloszlásával.

A statisztikát másra is fel lehet használni. Közismert, hogy Gárdonyi Géza titkosírásos üzenetet hagyott az utókor számára, melyet csak a 60-as években fejtettek meg. A megfejtő egy középiskolás diák volt, Gilicze Gábor, aki abból indult el, hogy a szöveget magyarul írták. Azt gyanította, hogy az írást a szokásos betűírással kódolták, csak a karakterek képe volt egyedi. Kis gondolkozással megtalálta a leggyakoribb magyar karaktert, a *t*-t, majd így haladt tovább. Más úton, de szintén abból a feltételezésből, hogy betűírásról van szó, a számokat – pontosabban a keltezést ismerte fel Gyürk Ottó alezredes. Ő számítástechnikában jártas felnőtt volt, és a hadseregben volt is hozzáférése számítógéphez, feltehetően Minszk 32-eshez, így a két megközelítésből kiindulva kezdték el dekódolni, és kitölteni a szöveget. Az egyszerű statisztika, ha nem is minden esetben, segített. Nem csak egyes karaktereket, hanem betűcsoportokat kellett kiértékelni.

A betűstatisztikának további gyakorlati jelentősége a tömör szövegtárolásnál van. Közismert, hogy az egyik egyszerű szövegtömörítési eljárás a Huffman kód. Ha közismert egy nyelv betűstatisztikája, akkor szövegfüggetlen – de nyelvfüggő – Huffman kódot lehet alkalmazni, tehát nincs szükség a segédtáblázat mellékeléséhez a szöveg továbbításához.

A nyelvészetben gyakran vizsgálják a több karakter egymás utáni előfordulását. Ezek sokkal jellemzőbbek egy nyelvre, mint az egyedi karakterek eloszlása. Egy orosz anyanyelvű nehezen tudja elsajátítani, hogy *i* előtt, *e* előtt a mássalhangzó lehet kemény is, így az *internet* kiejtésben gyakran *intyernyet*. Valószínűleg a honfoglaló magyarok nyelvében a szavak úgynevezett nyílt szótagú szavak voltak – nem torlódtak a mássalhangzók – ezért a régi időben nyelvünkbe került szavak honosodásakor magánhangzókkal bővültek. Pl. *kresztyán – keresztyén*, *dreko – derék*, *krcsma – korcsma*, *kricsmi*, de még a *gróf* szót is *gerófnak* ejtette sokáig a vidéki ember. Amikor egy francia vagy német angolul beszél, akkor általában a saját nyelvének hangzócsoportjai dominálnak, nem a briteké. A magyarokról nem is beszélek.

Ha arról készítünk statisztikát, hogy egyes karaktereket milyen valószínűséggel mi követi, akkor persze nagyobb szöveget kell elemezni, hogy megfelelő megbízhatóságú legyen mérésünk. Sőt, lehet trigráfokról, vagy általában n-gráfokról készíteni felmérést. Felvetődik, hogy meddig érdemes elmenni. Itt egyrészt az a kérdés, hogy mire akarjuk használni mérésünket, másrészt, tudunk-e elegendő adatot elemezni a megfelelő statisztikához – vagy egyáltalán javul-e a statisztikánk megbízhatósága további anyagok feldolgozásánál. A kérdés azért sarkalatos, mert a karakter-n-esek vizsgálati anyagmennyisége n-nel exponenciálisan nő. Első látásra a tárolandó információ is úgy tűnik exponenciális mennyisége az n-nek, de élő nyelveknél nem így van. Érdemes kipróbálni a gyakorlatban. Pl. négy mássalhangzónál több (szinte) sohasem fordul elő egy magyar szóban egymás után. Karakter már inkább: *bányá**szsztr**ájk*, de ritka.

A statisztika megbízhatóságánál lényeges a feldolgozott anyag mennyisége. Ha az egyszerű karakterstatisztikához megfelelő pár kilobyte, akkor a kettes statisztikához megabyte nagyságrendű anyag kell, hármas, négyes betűcsoportokat giga-, terrabájtnyi mennyiség adatait lehet jól kiértékelni. Mint korábban említettem, a statisztika egy nyelv írott szövegeiből von le tanulságokat, de ez – és emiatt a levont tanulság is – függ a szöveg eredetétől is. Ha a szöveg eredetétől függ az eredmény, akkor viszont feleslegesnek tűnik a nagy pontosság, emiatt a fent említett gigabájtos mennyiségre nincs szükség. A gyakorlatban jobb eredményt kapunk a nyelvről, ha úgynevezett jól kiegyensúlyozott mintát veszünk, tehát ügyelünk arra, hogy a szövegek

forrása kellően változatos legyen, annak az arányában, amilyen céllal akarjuk az eredményt hasznosítani. Az ötös és nagyobb betűcsoportoknál ekkor is érdemes lenne giga-, terrabájtokban gondolkozni.

Ha csak a statisztika áll rendelkezésünkre, akkor megpróbálhatunk ennek alapján az eloszlásnak megfelelő szöveget generálni. A generátor felfogható egy egyszerű stacionárius Markov-láncnak, tehát egy olyan véges automatával lehet reprezentálni, amelynek az állapotátmenetei valószínűségi alapon működnek. Az állapotok száma nagy n-nél magas lehet, de ha elegendő memóriánk van, akkor érdemes így reprezentálni. Ha nem, akkor könnyen lehet memóriában gazdaságos, de időben kellő sebességű algoritmust találni. Az generálási eredmény látványos. A nyelv Markov-lánc jellegét Shannon (1951) és Küpfmüller (1954) érdekes kísérletei is igazolják. Az alábbiakban Küpfmüller betűsorait mutatom be:

1 - *EME GKNEET ERS TITBL VTZEN*
2 - *AUSZ KEINU WONDINGLIN DUFRN*
3 - *PLANZEUNDGES PHIN INE UNDEN*
4 - *ICH FOLGEMASZIG BIS STEHEN DISPONIN*

Az első minta olyan betűsort ábrázol, ahol az egyedi előfordulási gyakoriságokat vették figyelembe, majd a betűkettősökét, betűhármasokét, betűnégyesekét. Jól látható, hogy az n növekedtével egyre biztosabbak lehetünk, hogy németül írták a szöveget. A hármas statisztika csak a nagyon közeli nyelveket nem különíti el. Az ötös statisztikával (annak idején ilyet már nem készítettek a technikai korlátok miatt) már olyan szövegek keletkeznek, melyre azt mondaná az anyanyelvű is, hogy nem értem, pedig a nyelvemen írták. Időnként értelmes mondatok is generálódhatnak. Az eredmény hasonló más nyelvekben is. Például magyar nyelven elvégzett hasonló, de kis és nagybetűket megkülönböztető elemzés után a következő generátumokat kaphatjuk a különböző hosszúságú statisztikákból:

1. *nórmlgenytnaeőettesy sdrúk üze neée issgnis k, őentetnsmrdttca*
2. *Bégiz érj, ercszépot, ekalfönembenyett kmáb os alőrre*
3. *Ezenlapoltike lehátszökmég hog, aztradon két. Korcsant ván.....*
4. *Hárone - Az a léhez, meg, akarózsák hogy öregész egény látán, járnagyon kérdeket egy emberet. Tordárd sem tan.*
5. *Senki adsaadott, hogy egy ki tornak, a kapta a mielőtte szor ismerni most vele aótsap a eipővel. - Ó, biztosítás úr csak társallatság vele. Hallj ott*
6. *Történik. Telt év óta meglátta bevallott kissé lehajtott és mindig légéz erligmyúneáenl eltűnt,*

Négyes statisztika után lényeges javulást nem látunk a generált halandzsán. Ennek az az oka, hogy a minta nem volt túl nagy. 7-8 megabájtnyi anyagon tanult a program, és a statisztika a négyes csoportok eloszlásának megbízhatósága már nem volt elegendő. Emiatt a generátum közel volt a determinisztikushoz, ami az n növelésével csak erősödött. Értelmesnek tűnő szövegdarabok keletkeznek, ami természetes, de ezek aránya nagyobb a vártnál, illetve annál, mint amikor esetleg giga- vagy terrabájtnyi anyagon tanult volna a rendszer.

Erre utal **Zipf törvénye** is. Ennek lényege, hogy ha bizonyos nyelvi objektumok száma elég nagy (pl. a szavak ilyenek), akkor sorrendezve őket előfordulásuk gyakorisága szerint, a relatív gyakoriság és a sorrendjük fordított arányban állnak. Pontosabban, $f_k=\alpha/k^s$
ahol f_k a k-adik elem relatív gyakorisága, s egy 1-hez közeli, annál valamivel nagyobb konstans, α pedig a normalizáló faktor, vagyis $\alpha\Sigma_k 1/k^s=1$. A törvény tapasztalati. Nem láttam még matematikai alátámasztását, de a gyakorlatban igaznak mutatkozik. A rangsorban elöljáró elemekre, tehát a leggyakoribbakra nem érvényes – itt valamilyen hatás miatt általában kisebb a gyakoriság, mint az feltételezhető lenne. A sor végén szereplőknél más miatt nem lehet mérvadó a számítás. Ha pl. a szövegekben előforduló szavak nézzük, akkor a gyakoriak a szöveg nagy részét lefedik, de a rangsor második harmadában szereplő szavak már egyszer, legfeljebb kétszer fordul elő a szövegben. Gyakoriságuk, mivel csak egész érték lehet, egy vagy legfeljebb kettő. A valószínűség alapján viszont többnyire történtékű lehetne az előfordulásuk, tehát elég nagy relatív hibával becsülhető a valószínűségük. A legjobban a középmezőny elemezhető Zipf törvénye alapján.

Ami a karakterstatisztikát illeti, betűírás esetén az egyes karaktereloszlásnál nem alkalmazható a törvény, mert – bár sokszor fordulnak elő – de kevés a megkülönböztethető egyed. Más a helyzet az ideografikus jeleknél, vagy ha 3-, 4-gráfok eloszlását vizsgáljuk.

Érdemes még azt vizsgálni, hogy milyen adatstruktúrát érdemes használni a betű-, karakterstatisztikáknál. Nyilvánvaló, hogy a szimpla, a kettes, esetleg a hármas gyakoriságmérésre megfelelhet az egy, két illetve háromdimenziós tömb, melynek indexei az n-edik karakter kódja. 8 bites ábrázolásnál ez kb. 10 000 elemű tömb. Ez célszerű, és gyors elérést tesz lehetővé, hisz egy indexelés után közvetlen elérhetjük a kívánt elem számlálóját. Ha viszont megnézzük Zipf törvényét, vagy a gyakorlatban számolni kezdjük az előfordulásokat,

akkor kiderül, hogy a legtöbb elem egyszer sem fog szerepelni, vagyis a gyakorisági mátrix rendkívül ritka. Ezért ajánlatos más reprezentációt választani – vagy a szokásos ritka mátrixok módszerét kell használni, vagy egyéb trükköt kell bevetni. E nélkül már az 5-ös karakterstatisztika mátrixa oly nagy, hogyha a gépen reprezentálható is, valószínűleg virtuális memória használatára kényszerülne a program, ami megsokszorozhatja a futásidőt.

Az ilyen jellegű statisztikai módszereknek nagy a gyakorlati haszna. Erre később még visszatérek.

Feladat 1: Készítsen programot betűstatisztikára, és ennek alapján elemezzen különböző élő és programnyelven írott szövegeket. A gyűjtés alapján generáljon szövegeket 1-es, 2-es, 3-as, esetleg nagyobb betűcsoportok adatainak felhasználásával. Értékelje ki az eredményt.

Feladat 2: Tegye sorrendbe a karaktereket gyakoriságuk szerint. Nézze meg az eloszlás görbéjét. Vesse össze az ilyen eloszlást egy írótól származó különböző korpuszon készített statisztikán, és különböző szerzők műveiből mért eloszlásokat.

Feladat 3: Vizsgálja kétszer logaritmikus skálán a kínai írásjelek gyakoriságát. Becsülhető-e az összes kandzsi száma a statisztika alapján?

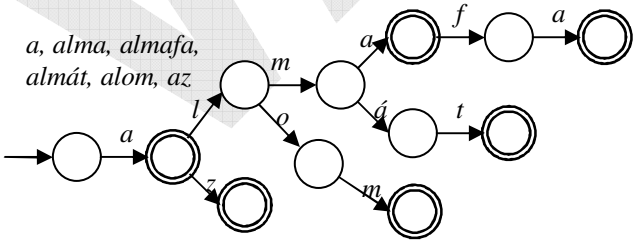
Feladat 4: Elemezze különböző nyelvek 3-grammjainak gyakoriságát. Lehet-e becsülni, melyik nyelv sűrűbb, vagyis mely nyelvekben gyakoribb, hogy két szó alig különbözik?

3.8 A sorba rendezés tudománya.

A keresési algoritmusok alapfeltétele a jó rendezés.
(Donald Knuth)

A betűírás egyik nagy előnye, hogy a betűk sorba rendezésével egyszersmind a szavak, szövegek sorba rendezése is megoldott. A szokásos ábécébe rendezés matematikailag is jól megfogalmazott algoritmus, aminek alapján telefonkönyvben, szótárban könnyedén találjuk meg a keresett nevet, címszót. A nyelvészeti programok is mindig óriási mennyiségű halmazokkal dolgoznak, amelyben a legfontosabb, hogy gyorsan megtaláljunk bármit, illetve eldöntsük, egy keresett objektum eleme-e a halmaznak, vagy sem. A sorba rendezésnek alapvető célja, hogy a keresést meggyorsítsa. Keres a telefonálni akaró a telefonkönyvben, a fordító a szótárban, de a gépi adatfeldolgozás egyik kulcsa a jó rendezés, keresés.

Rendezésre nagyon sok jó algoritmus létezik. Ezekből kiemelném a kiegyensúlyozott fákat, amelyek segítségével a keresést (látszólagosan) logaritmikus nagyságrendben hajtják végre, legtöbb esetben a beszúrás és törlés is ilyen sebességgel elvégezhető. A valóságban ez azért nem igaz, mert az összehasonlítás művelet nem egység idejű, ez függ a bemenet hosszától. Tehát, ha n a már tárolt szavak száma, és m a keresendő szó hossza, akkor a keresési idő kiegyensúlyozott fák esetén $\mathrm{Ord}(m^*\log(n))$. Ha nyelvi információkról van szó, akkor ez egy lényeges tényező. Ebben az értelemben, a keresés legalábbis a keresendő objektum (karakterlánc) hosszával sem arányos, annál nagyobb. Létezik olyan struktúra, ami ezt a nagyságrendet nem lépi túl. Ez a betűfa (más néven szófa, vagy angolul trie). A szófában karakterenként csomópontról csomópontra haladva azonosíthatjuk, hogy az eddig olvasott prefixnek van-e útja a fa gyökerétől. Ha a karakterlánc végére értünk, csak azt kell ellenőrizni, elfogadó csomópontba értünk vagy sem. Ezért a keresés ideje $\mathrm{Ord}(m)$.



Könnyen látható, hogy a beszúrás és a törlés is az input hosszával arányos időben elvégezhető, tehát elméletileg a lehető leggyorsabb. A gyakorlatban a konstans is kellően kicsi, hisz az egyes lépések rendkívül elemiek, számítógéppel könnyen megvalósíthatók. A lépésenkénti összehasonlítást az aktuális karakterek kódja alapján lehet elvégezni.

Az adattárolás szempontjából is kedvező a betűfák alkalmazása. A különböző kiegyensúlyozott keresési fák esetén (red-black tree, b-tree, kupacok…) az objektumokat, vagyis a szavakat általában teljes egészében kell tárolni a csomópontokban, míg a szófák esetén a közös prefixek minden karaktere csak egy helyen szerepel függetlenül attól, hány szónak szerepel az elején. A nagy trükk alapvetően az, hogy olvasás közben karakterenként döntünk a keresőfában való lépesről is, ami az abc-rendezésnek megfelel. Ha továbbléptünk, már nem érdekel a szöveg korábbi része.

Az Egyszerű ábécébe rendezés sajnos nem mindig jó a természetes nyelveknél.

1. Az egyik gond lehet, hogy a karakterek kódja nem mindig az ábécé sorrendjét követik. Az ékezetes betűk a legtöbb nyelvben az éktelen párjukat közvetlen követi az ábécében, de kódjuk az összes éktelent

követi. A kód és a sorrend különbségét, átkódolással, súlyfüggvénnyel könnyen korrigálhatjuk. Ezzel az is megoldható, ha két különböző betűnek azonos a súlya, hisz lehetnek olyan betűk, melyeknek azonos a pozíciójuk az ábécében.

2. Sok nyelvben – a magyarhoz hasonlóan – a betűknek van elsődleges és másodlagos sorrendjük. A rövid és a hosszú magánhangzók közti különbség nem befolyásolja két szó összehasonlítását, ha van olyan karakter, amely eldönti, melyiket kell előbbre venni. Ha ennek alapján azonosnak mutatkoznak a szövegek, akkor kell megvizsgálni, rövid, vagy hosszú volt-e a magánhangzó.

   ver, vér, veréb, véreb, verebek, veres, verés, véres…

   a helyes sorrend. Meg kell említenem, hogy a franciában a másodlagos rendezés nem balról jobbra, hanem jobbról balra történik, ezért a fenti szavakat (ha francia szavak volnának) a következőképpen kellene rendezni: *ver, vér, véreb, veréb, verebek, veres, véres, verés…*
   Az egyszerű súlyozás helyett itt kettős súlyozás megoldja a gondot.
3. A legnagyobb gondot az okozza, hogy az akadémiai szabályok szerint a sorba rendezés nem karakterre, hanem betűre vonatkozik. Persze ha a súlyfüggvényt betűnként definiáljuk, akkor megoldottnak tekinthetnénk a feladatot. A gond ott van, hogy a szavakban meg kell találni azokat a karaktercsoportokat, amelyek egy-egy betűt jelölnek. A *malacság* és a *kocsány* szóban ugyan az a két karakter egyszer egy, másik esetben két külön betűt jelöl. Ha a *malachit* szót vizsgáljuk, akkor a *ch* párt nézhetjük egy betűnek, amikor egy márványszerű anyagra gondolunk, de – egy kis szabad fantáziával – lehet két külön betű, ha sertések vallására gondolunk. Megoldás lehet, ha a szavakat elemezni tudjuk, mert akkor az esetek döntő többségében egyértelművé válik az összes karaktercsoport milyensége. A bajt az is súlyosbítja, hogy a Magyar Szabvány alapján pl. a *ch* két betűnek tekintendő a sorrendezésben, kivéve a történelmi neveknél, ahol a kiejtés szerinti *h* betűt követi. Emiatt a szabvány és a sorrendezés elve ütközik a *technika*, *mechanika* szavainknál, ami kissé érthetetlen.

A technikai nehézségek miatt többféle sorrendezés létezik

1. Akadémiai – betartja az összes szabályt. Ezt kell alkalmazni szótáraknál, lexikonok szerkesztésénél. Nyelvi háttértudás nélkül megvalósíthatatlan, illetve csak jelentős hibaszázalékkal lehet megvalósítani.
2. Technikai – számítógéppel jól kivitelezhető, ezért nem foglakozik a digráfokkal, trigráfokkal. Gyakran változó, emberi olvasásra szánt anyagoknál használják, pl. telefonkönyvek készítésénél.
3. Gépi célú – a lényeg az, hogy a keresés, befűzés, törlés gyors legyen, ezért tetszőleges, a célnak megfelelő egyszeres súlyfüggvényt használ, ha egyáltalán használ súlyfüggvényt. Nem emberek, csak a gép számára sorrendez.

A szótárak elemeit az Akadémiai módszerrel illik rendezni. Ez emberi használatra teljesen jó, ha ismert a nyelv a felhasználó oldaláról. Az idegennek gondot jelenthet a keresésben a két-háromjegyű karakterek felismerése. Ennél nagyobb gond, ha a szövegben azt látja, hogy *terhet*, és ki akarja keresni a szótárban, mit is jelenthet, akkor már a harmadik betűnél eltéved, mert természetesen a *teher* szócikknél kéne keresnie. Ez egy idegen számára nem magától értetődő.
Ennél nehezebb helyzetben van, aki héber, arab szótárat használ. Nem a balról jobbra írás a gond, hanem az, hogy a sémi nyelvekben a szónak elő és utótagjai vannak. Emiatt a ragozott szóalakhoz tartozó szócikket nem karakterenkénti abc sorrendben találja meg. Fel kell ismerni, mi lehet a szó töve, és ott kell keresni. Pontosabban, ezekben a szótárakban a szócikkek szógyökökként vannak összegyűjtve. A szógyök sok szónak a közös része. Így a hatalom, terület, birtokol stb. mind egy gyökre vezethető vissza, ezért ugyanabban a tartományban kell keresni.
Még cifrább az ideografikus írás esete. Ugyan vannak szavak, melyeket több karakterből tesznek össze, de a karakterek sorrendje nem megjegyezhető, mint a betűírás esetén. Emiatt a kínai írásban a karakterek sorrendjét a benne levő íráselemek határozzák meg. Azok sorrendje, melyekből persze nincs több ezer, már megtanulható.

Feladat 1: Készítsen gyors többsúlyos könnyen paraméterezhető rendezőprogramot!

Feladat 2: Készítsen olyan rendezőprogramot, amely a digráfok (trigráfok) rendezését is megoldja – interaktív, tanuló módon vagy valószínűségi alapon!

4. Szó(alak)tan

Amikor osztályozunk valamit, jó, ha úgy tesszük, hogy megjegyezhető, jól kezelhető osztályokat használjunk. A gépnek ugyan mindegy, de az embernek nem kellemes, ha egy szempont szerinti osztályok száma magas. A kognitív pszichológiának van egy bűvszáma, az öt. Nem pontosan öt, lehet kevesebb vagy kicsit több, nyolc, esetleg tíz, de ennél több nehezen tanulható, áttekinthető. Ha valamit sokfelé kell osztani, lehetőleg állapítsunk meg olyan független szempontokat, melyek szerinti osztályozás már nem lépi túl a megérthetőség korlátját.

4.1 Mi a szó?

Pontosabban mi egy nyelv szavainak halmaza? A szó beszédünk legkisebb értelmes része – tanultuk általános iskolában. Hát ez nem igaz. Tudjuk, egyes szavak önmagukban nem állnak meg (pl. névelő), más szavak értelmes részekre darabolhatók (pl. összetett szavak). Mégis tudjuk, mi a szó. Írásban szóköz nélkül egybe írt része a szövegnek. De nem tetszőleges. Papp Ferenc is ezt a kérdést teszi fel. Az-e a szó, amit egy nyelven leírtak, vagy akár leírhattak volna? Már az alapvető kérdésre sem egyszerű a válasz: Mi a szó? A kérdésnek két aspektusa van. Egyrészt, mit fogad el a nyelvtudomány helyesen alkotott szónak, vagy mit használ a magyarul beszélő. Másrészt a számítástechnika szempontjából a szöveget mindenképpen fel kell darabolni szavakra, hogy a további elemzéseket megtehessük.

Emiatt első közelítésben nézhetjük a szöveg szóköz, tabulátor, sorvég közti részeit, és ezt tekintsük szónak. Persze vannak írásjelek is, számok és egyéb karakterek, melyek közt egyesek szóalkotó karakterek lehetnek, mások nem. Pl. a *?* , *!* , *:* , *;* nem szóalkotó karakter, de a *'* előfordul egyes szavainkban. A . lehet része a szónak, ha egy rövidítés végén áll. A - szóalkotó bizonyos szóösszetételeknél, toldalékoknál, de máskor nem. Ha a számjeggyel írt számok is szavak, akkor a , lehet a szám belsejében tizedesvessző. Más nyelvben a : is lehet része a szónak.

A programnyelvek elemzésénél a lexikális analízis alatt a szavak elkülönítése egyszerű, mert formális módszerrel definiálták. Az élő nyelvi elemzésnek az első lépése is a lexikális analízis, de az informális definíció formalizálása közben nem mindig egyértelmű a művelet.

4.2 Hány szó van egy nyelvben?

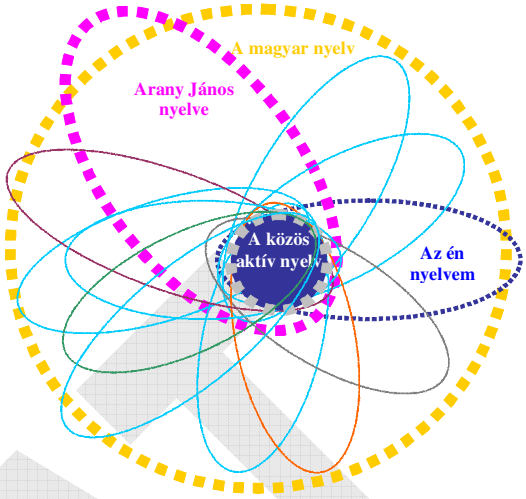
Már az alapvető kérdésre sem magától érthető a válasz: Mi a szó? A kérdésnek két aspektusa van. Egyrészt, mi nevezhető különböző szónak, másrészt mi nevezhető magyar szónak.

Az első szempontot demonstrálandó mindenki külön szónak tekinti azt, hogy *kutya, macska, eszik…* Ezzel szemben azonos szó alapvetően a *kutya, kutyát, kutyáimról, kutyához…* Ezek egy szónak **különböző alak**jai. Fogalmilag jó határozottan különválasztani a szót, a lehetséges alakjaitól. Az előzőt általában **lexémá**nak, esetleg **lemmá**nak, **szótári alak**nak nevezik, míg az utóbbit **szóalak**nak, **szóformá**nak. A szó megjelölést mindkét értelemben használják.

A másik szempont szerint az a kérdés, hogy egy nyelv szavának egyáltalában mi tekinthető. És itt most nem arra gondolok, hogy a nyelv időben változik. Természetesen az is a tudomány vizsgálatának tárgya, hogy hogyan változik egy nyelv, jelen témát figyelembe véve egy nyelvben mikor jelenik meg egy új szó, mikor kopik ki a használatból. Ha ettől eltekintünk, akkor is feltehetjük a kérdést, mi tekinthető mai magyar szónak. Azt nem mondhatjuk, hogy az, amit minden magyar anyanyelvű beszédében, írásában használ. Ha felmérést készítenénk, meglepően kis szókészlet jönne ki. Egy jó tollú újságíró egész életében mintegy tízezer különböző szót használ. Egy hétköznapi állampolgár ennek kétharmadát. A felmérések szerint, legtöbb magyar szót Jókai Mór vetett papírra, mintegy harmincezret. Ő persze gyakran használt rövid életű divatszavakat, amelyeket ma már meg sem értenénk. Másodiknak Arany Jánost tartják huszonötezer szavával, Petőfi Sándor rövid élete alatt a húszezret is alig érte el. Ha viszont azt nézzük, ezen írók, költők szókészletében mennyi a közös, akkor nem hiszem, hogy hatezer szónál többet találnánk, pedig nagyjából ugyanannak a kornak a tollforgatói voltak.

Adott nyelvnek a szókészlete ennek ellenére becsülhető. Legtöbb nemzetnek létezik szógyűjteménye. Ilyenek a kétnyelvű szótárakból, lexikonokból, értelmező szótárból kigyűjthetők. A magyar nyelvi munkák többségének hivatkozási alapja a Magyar Értelmező Kéziszótár. Ez a gyűjtemény mintegy nyolcvanezer címszót tartalmaz. Hogy menyire fedi le a nyelvet, azon lehet vitatkozni, mindenesetre tartalmazza az összes gyakran használt szavunkat. Vannak benne ritkán használtak is (*kardalésza, suly, süly, garádics…*), melyek egy része elavult, más részét csak egyes táján használják az országnak. Ritka az olyan szó, melyet ma is országszerte használnak, de nincs benne, mert kimaradt (*kozmál, csevej*). Szakszavak kevésbé szerepelnek benne, tulajdonnevek pedig csak esetlegesen, hiszen ez egy értelmező szótár. A szótár mérete ennek ellenére mérvadó.

A szavak többségét minden magyar anyanyelvű érti. Ha találkozik vele szövegben, semmi gond sem merül fel értelmezésével. Más nyelv hasonló szógyűjteménye hasonló méretű. Legyen az angol, orosz, cseh, spanyol. Valahol 50 000 és 300 000 tétel szerepel egy-egy nép nyelvi szókincstárában.



Az a szómennyiség, amit az egyes emberek használnak, ennek a mennyiségnek a tizede. Ha valaki elsajátít 6000-10000 szót egy idegen nyelven, akkor az már jó szókincsnek számít, de akkor sem fog használni 1000-2000 szónál többet a tanult nyelven. Ennek alapján meg lehet különböztetni egy **aktív** és egy **passzív** szókészletet. Ha az ember idegen nyelvet tanul, akkor természetesen merül fel a két fogalom megkülönböztetése, de anyanyelvi téren ugyanúgy létezik.

Függetlenül a szókészlettől, a nyelv mégiscsak a nyelvet beszélők összessége határozza meg. Más az én magyarom, mint Józsi bácsié, a sarki fűszeresé, Torgyán Józsefé, a vonatkalauzé. Ezekből tevődik össze a magyar nyelv. Nem additívan, tehát nem úgy, hogy mindenki belerakja a magáét, hanem úgy, hogy amit többen használnak, értenek, az része a nyelvnek akkor is, ha egyesek esetleg nem értik.

Persze a nyelv szókincse nem egyszerűen a nyelvet beszélők szókincsének összessége. Ha a ragozott alakokat nem is tekintjük külön szónak – a szótárban nem szerepelnek független tételként – a képzett alakoknál már más a helyzet. Olyan esetben, ha a képzett alak jelentése, fordítása más nyelvre nem közvetlen következik a képzés módjától, akkor a képzett alakok a szótárakban függetlenül szerepelnek. Annak a megítélése, hogy a képzési szabály mennyire reguláris, szubjektív. Angol nyelvű szótárak gyakrabban teszik külön szócikkbe a képzett alakokat, mint a magyarok.

Ha az ábrát úgy értelmezzük, hogy a középpontban vannak a nyelv gyakrabban használt elemei és a periférián a ritkán használatosak, akkor jól látszik, hogy van egy ezerre tehető készlet, amely közös, és százezrekben mérhető egyéb szó. Néha ki-kilóg a nyelvből egyes emberek nyelvhasználata. A szóösszetétel is produktív a magyar nyelvben. Lehetetlen felsorolni a lehetséges szóösszetételeket, ezért csak azok szerepelnek a szótárakban, amelyeknek valamilyen módon különös, nem levezethető jelentése van, vagy fordítása egy másik nyelvre jellegzetes (*kékfestő*, *osztályvezető*…). Ebből a szempontból a német nyelv hasonló a magyarhoz, sőt ott sokkal gyakoribb a szóösszetétel. Az angolban ritka eset, amikor egybeírják, esetleg kötőjellel kapcsolják a szavakat, ezért náluk minimálisan növeli a szótár terjedelmét, ha mindegyiket külön tételnek veszik fel.

Ha a magyar szókincset – például a helyesírási kézikönyv szavait – vizsgáljuk, akkor kiderül, hogy a mintegy 75 000 szó előállítható 30 000 szó szabályos képzésével, illetve szóösszetételeként. Akkor most ezek az előállított szavak önálló elemei a nyelvnek, vagy kisebb a szókészletünk? Erre a válasz az, hogy attól függ, mire használjuk szótárunkat. Ha a fő kérdés, hogy minek mi a jelentése, mivé kell fordítani, akkor meg kell különböztetni a *házasság* szót a *ház*-ból képzett szótól. Ha csak a szavak formájával foglalkozunk, akkor – modelltől függően – lehet, hogy elegendő a *ház* szóval foglalkozni. Más nyelvekben is hasonló gondok vannak, de a nagyságrendi becslések, mely szerint egy nyelv szavainak száma 500 000 körül van nyelvfüggetlennek mondható.

4.3 Hány alakja van egy szónak?

Az előző fejezet alapján felmerül a kérdés, lehet, hogy szóalakból is körülbelül ugyanannyi van minden nyelvben? Erre a választ egy kis számolással adjuk meg.

Vegyük az angol nyelvet. Egyes mellékneveknek van közép- és felsőfoka. Ha a *-ly* képzővel határozót is lehet formálni a szóból, de így sem jöhet ki 4-nél több alak (*easy, easier, easiest, easily*). Ha még hozzávesszük a nem általánosan képzett alakokat (*easiness, uneasy, uneasily*), akkor is 7-nél maradunk. A főneveknek egyes és többes számú alakjuk van, ezeket esetleg birokos jellel cifrázhatjuk, de így is csak 4 formában jelenhetnek meg a szövegben (*table, tables, table's, tables'*). Az igék ragozása is szegényes. A létigének (*be*) és a birtoklás (*have*) igének is **10** alatt van a ragozott formája (*have, has, had, having, hasn't, haven't, hadn't*).

A latin nyelvek szótana (latin, spanyol, olasz, francia, román…) ennél összetettebb. A ragozott ige nem csak a cselekvő számát, személyét fejezi ki, de az igemód, igeidő is tükröződik a ragozásban. Emiatt egy szónak akár megközelítően **100** (ritka modellekben egyeseknek száznál több) alakja lehetséges.

A szláv nyelvek már igazán próbára teszik a nyelvtanulókat. A melléknevek ragozásában kifejeződik a jelzett neme (3-féle) száma (egyes, többes) és az esetrag (6-féle). Létezik rövid és hosszú alakjuk, és határozó is képezhető belőlük. Ezzel persze még az ötvenet sem éri el az alakok száma, de ha veszünk egy átlagos igét, akkor az igeragokon kívül (folyamatos-befejezett, cselekvő száma, személye, múlt idő, visszaható alakok, összesen 40 alak), minden igéből lehet határozót és melléknevet képezni, és a mellékneveknek újabb toldalékolt formái vannak. Így határozottan százakban mérhető az esetlegesen egy szóhoz tartozó alakok száma, és ekkor még nem vettük számításba az igekötőt, a melléknevek esetleges közép- és felsőfokú alakjait, a tagadó alakokat. Emiatt nem rossz becslés, ha azt állítjuk, hogy az oroszban – általában a szláv nyelvekben – egy szónak közel **1000**, esetleg több alakja lehet.

Lehet ezt fokozni? Lehet, de a pontosabb számolást későbbi fejezetekre hagyom. Egy kis becslést azért tehetünk. A magyarban minden névszó kaphat ragokat meg jeleket. A lehetséges ragok száma – modelltől függően – 17-30 lehet. A jelek: többes szám, birtok- és birtokos jel ettől függetlenül 14 lehetőséget ad, nem beszélve a családi többesről. Ez mintegy 600 lehetőség. Az igerag kifejezi a cselekvő számát, személyét, és 5 ragozási mód van. Ez az érték sem sok, 60 körüli. A nagy szám nem innen származik, hanem a képzőkből. Hiszen (majdnem) minden névszóból képezhetünk igét (*-oz, -ít, -odik*…), igéből folyamatos vagy befejezett melléknévi igenevet, vagy főnevet (*-ó, -atlan, -ott, -ás*), de képzők segítségével igéből képezhetünk újabb igét (*-gat, -tat, -hat*), főnéből melléknevet, melléknévből főnévet (*-os, -ság*). Ha minden szóhoz csak négy különböző képző kapcsolódhat, és ezt a folyamatot háromszor megismételhetjük, akkor egy szónak $4^3+4^2+4^1+4^0=88$ képzett alakját kapjuk, ami ha névszó, 600 ragozott, jelzett alakja lehet. Ez önmagában több mint **10 000**.

Egy kis táblázatban azért összefoglalnám az adatokat, melyben nem csak az egyes szavak alakjainak számát, hanem a nyelv összes szóalakjának számát is becsülöm. A felülről becslés alapja egy 100 000 szóból álló alapszótár, az alulról való becslésé egy 20 000 szóból álló szótárt feltételez. Ha ezt megszorozzuk az egy szóhoz tartozó alakok számával, megkapjuk az utolsó sor adatait.

nyelvcsalád	Angolszász nyelvek	Latin nyelvek	Sémi nyelvek	Szláv nyelvek	Finnugor, török nyelvek
egy szó használatos alakjainak száma	< 5	< 100	< 200	< 1000	> 10 000
egyes modellekben maximum	10	200	400	2500	korlátlan?
az összes szóalak száma	< 400 000	< 3 000 000	< 6 000 000	< 50 000 000	> 200 000 000

4.4 Szóstatisztikák

A tapasztalatok szerint egy sok éven át finomodó nyelvi adatbázis 10 000-100 000 elemnél válik javíthatatlanná. A számot az adatbázis építésének módja és a tárolt elemekre vonatkozó Zipf konstansok határozzák meg. Emiatt nem egyértelmű, hogy a korábbi kiadáshoz képest ténylegesen javult-e az Akadémiai Kiadó Angol-Magyar nagyszótára. Hogy bővült, az biztos, de a bejegyzések megbízhatósága nem lett jobb.

A betűstatisztikához hasonlóan szóstatisztikát is lehet készíteni. ebben az esetben persze meg kell különböztetni a szóformát a lemmától. Ha szóformákra készítünk statisztikát, akkor is kérdéses, a kis- és nagybetűket megkülönböztetjük-e. Ha megkülönböztetjük, minek számít a mondat eleji karakter.

Ha a szavak lemmáinak gyakorisága érdekel, akkor a legnagyobb gond a többértelműség, Ha ezt valamilyen biztonsággal megoldjuk, akkor számíthatunk megbízható eredményekre. Ha egy ilyen statisztikát nézünk, akkor szembetűnő, hogy a leggyakoribb szavak előfordulása igen magas, míg többségben lesznek az egyszer-kétszer előfordulók.

A Zipf törvény itt igazán alkalmazható. Maga Zipf is szavak gyakoriságának vizsgálata közben jött rá a törvényszerűségre. Nézzük például a kínai írás karakterstatisztikáját. Az első oszlop a karakter gyakorisági sorrendjét mutatja, a második a karaktert, a harmadik pedig azt, hogy az első n karakter a teljes szöveg hanyadrészét fedi le:

1	的	3.2360%	77	北	34.3432%	153	点	48.5222%	229	城	58.2062%
2	国	4.6229%	78	下	34.5668%	154	及	48.6698%	230	赛	58.3159%
3	人	5.6785%	79	产	34.7883%	155	表	48.8174%	231	增	58.4255%
4	一	6.7305%	80	实	35.0092%	156	队	48.9648%	232	口	58.5347%
5	中	7.7317%	81	动	35.2297%	157	计	49.1120%	233	费	58.6437%
6	学	8.6756%	82	重	35.4491%	158	品	49.2578%	234	式	58.7524%
7	在	9.5410%	83	化	35.6667%	159	总	49.4032%	235	题	58.8607%
8	大	10.3130%	84	可	35.8841%	160	期	49.5474%	236	手	58.9686%
9	有	11.0797%	85	位	36.1011%	161	好	49.6903%	237	安	59.0765%
10	年	11.8347%	86	得	36.3174%	162	委	49.8318%	238	因	59.1843%
11	了	12.5689%	87	同	36.5327%	163	都	49.9719%	239	情	59.2892%
12	是	13.2565%	88	京	36.7478%	164	回	50.1109%	240	管	59.3941%
13	和	13.8576%	89	区	36.9617%	165	里	50.2493%	241	共	59.4982%
14	不	14.4221%	90	将	37.1700%	166	两	50.3866%	242	组	59.6022%
15	为	14.9658%	91	资	37.3781%	167	并	50.5238%	243	统	59.7060%
16	上	15.5043%	92	过	37.5847%	168	项	50.6609%	244	十	59.8095%
17	会	15.9892%	93	名	37.7904%	169	提	50.7974%	245	女	59.9119%
18	家	16.4600%	94	之	37.9958%	170	由	50.9338%	246	省	60.0142%
19	生	16.9236%	95	万	38.1997%	171	特	51.0692%	247	请	60.1164%
20	业	17.3772%	96	机	38.4025%	172	济	51.2040%	248	商	60.2184%
21	地	17.8295%	97	也	38.6049%	173	程	51.3378%	249	先	60.3196%
22	出	18.2591%	98	内	38.8069%	174	企	51.4716%	250	相	60.4202%
23	个	18.6650%	99	所	39.0072%	175	界	51.6052%	251	处	60.5207%
24	工	19.0705%	100	关	39.2073%	176	香	51.7383%	252	编	60.6206%
25	这	19.4676%	101	就	39.4074%	177	着	51.8712%	253	流	60.7202%
26	以	19.8601%	102	技	39.6072%	178	利	52.0040%	254	领	60.8199%
27	成	20.2465%	103	能	39.8067%	179	度	52.1367%	255	话	60.9189%
28	发	20.6319%	104	已	40.0045%	180	使	52.2692%	256	别	61.0178%
29	作	21.0172%	105	加	40.2015%	181	当	52.4008%	257	改	61.1162%
30	我	21.4006%	106	元	40.3960%	182	各	52.5316%	258	运	61.2144%
31	日	21.7717%	107	者	40.5900%	183	华	52.6622%	259	道	61.3125%
32	来	22.1267%	108	留	40.7838%	184	向	52.7929%	260	周	61.4100%
33	科	22.4794%	109	法	40.9776%	185	专	52.9231%	261	每	61.5073%
34	行	22.8301%	110	定	41.1710%	186	今	53.0528%	262	件	61.6038%
35	到	23.1777%	111	说	41.3632%	187	立	53.1823%	263	无	61.6999%
36	市	23.5162%	112	合	41.5553%	188	去	53.3114%	264	取	61.7956%
37	多	23.8460%	113	术	41.7468%	189	书	53.4400%	265	二	61.8911%
38	要	24.1708%	114	体	41.9374%	190	联	53.5677%	266	或	61.9865%
39	时	24.4904%	115	次	42.1273%	191	数	53.6946%	267	但	62.0817%
40	经	24.8069%	116	小	42.3164%	192	些	53.8214%	268	单	62.1767%
41	高	25.1165%	117	展	42.5053%	193	如	53.9473%	269	山	62.2716%
42	外	25.4003%	118	面	42.6929%	194	门	54.0730%	270	老	62.3661%
43	员	25.6840%	119	平	42.8788%	195	正	54.1981%	271	议	62.4599%
44	公	25.9651%	120	院	43.0627%	196	士	54.3225%	272	据	62.5521%
45	对	26.2421%	121	其	43.2460%	197	达	54.4464%	273	结	62.6438%
46	海	26.5170%	122	政	43.4271%	198	南	54.5702%	274	持	62.7353%
47	部	26.7914%	123	心	43.6080%	199	活	54.6938%	275	原	62.8267%
48	们	27.0635%	124	事	43.7882%	200	量	54.8164%	276	投	62.9173%
49	分	27.3344%	125	而	43.9680%	201	办	54.9385%	277	果	63.0078%
50	进	27.6049%	126	金	44.1475%	202	广	55.0605%	278	被	63.0975%
51	开	27.8738%	127	场	44.3262%	203	育	55.1824%	279	强	63.1866%
52	用	28.1423%	128	天	44.5043%	204	然	55.3041%	280	受	63.2755%
53	子	28.4106%	129	第	44.6780%	205	车	55.4255%	281	服	63.3643%
54	月	28.6763%	130	力	44.8514%	206	东	55.5466%	282	农	63.4519%
55	全	28.9408%	131	水	45.0246%	207	交	55.6677%	283	规	63.5396%
56	方	29.2017%	132	物	45.1965%	208	此	55.7880%	284	看	63.6271%
57	民	29.4611%	133	究	45.3680%	209	间	55.9071%	285	才	63.7143%
58	等	29.7178%	134	最	45.5366%	210	西	56.0257%	286	台	63.8005%
59	于	29.9744%	135	入	45.7030%	211	近	56.1443%	287	少	63.8859%
60	目	30.2292%	136	设	45.8672%	212	路	56.2625%	288	常	63.9709%
61	新	30.4836%	137	务	46.0313%	213	任	56.3805%	289	获	64.0556%
62	他	30.7359%	138	制	46.1953%	214	至	56.4985%	290	选	64.1395%
63	本	30.9857%	139	三	46.3594%	215	校	56.6163%	291	头	64.2223%
64	长	31.2349%	140	明	46.5230%	216	保	56.7338%	292	又	64.3048%
65	现	31.4831%	141	美	46.6817%	217	社	56.8498%	293	放	64.3872%
66	自	31.7304%	142	还	46.8400%	218	性	56.9656%	294	只	64.4690%
67	后	31.9759%	143	从	46.9979%	219	应	57.0805%	295	集	64.5501%
68	教	32.2193%	144	代	47.1533%	220	收	57.1953%	296	神	64.6311%
69	建	32.4604%	145	通	47.3086%	221	司	57.3091%	297	职	64.7121%
70	文	32.6995%	146	基	47.4632%	222	际	57.4229%	298	局	64.7931%
71	理	32.9384%	147	港	47.6177%	223	州	57.5363%	299	光	64.8738%
72	研	33.1768%	148	世	47.7714%	224	问	57.6493%	300	参	64.9545%
73	前	33.4152%	149	系	47.9246%	225	江	57.7618%	301	条	65.0350%
74	主	33.6506%	150	种	48.0770%	226	信	57.8733%	302	几	65.1156%
75	电	33.8852%	151	起	48.2262%	227	意	57.9846%	303	举	65.1961%
76	与	34.1144%	152	比	48.3742%	228	报	58.0957%	304	团	65.2766%

305	接	65.3569%	381	支	70.8174%	457	布	75.2910%	533	县	78.9123%
306	调	65.4365%	382	息	70.8828%	458	演	75.3435%	534	岁	78.9558%
307	决	65.5158%	383	清	70.9481%	459	推	75.3959%	535	半	78.9989%
308	青	65.5951%	384	风	71.0131%	460	客	75.4484%	536	八	79.0419%
309	米	65.6740%	385	么	71.0777%	461	授	75.5007%	537	值	79.0849%
310	导	65.7526%	386	节	71.1422%	462	字	75.5527%	538	易	79.1277%
311	五	65.8311%	387	干	71.2066%	463	馆	75.6047%	539	太	79.1703%
312	四	65.9095%	388	治	71.2709%	464	款	75.6566%	540	责	79.2129%
313	求	65.9878%	389	府	71.3347%	465	整	75.7085%	541	钱	79.2555%
314	军	66.0662%	390	证	71.3983%	466	查	75.7602%	542	往	79.2980%
315	记	66.1444%	391	观	71.4620%	467	乡	75.8117%	543	尔	79.3405%
316	首	66.2226%	392	助	71.5254%	468	连	75.8629%	544	席	79.3830%
317	传	66.3006%	393	境	71.5887%	469	古	75.9138%	545	艺	79.4254%
318	解	66.3782%	394	难	71.6514%	470	均	75.9647%	546	评	79.4675%
319	更	66.4554%	395	王	71.7138%	471	河	76.0155%	547	仪	79.5096%
320	很	66.5325%	396	花	71.7762%	472	谈	76.0660%	548	申	79.5516%
321	房	66.6091%	397	革	71.8383%	473	转	76.1164%	549	斯	79.5935%
322	刊	66.6855%	398	批	71.9003%	474	足	76.1667%	550	初	79.6351%
323	亿	66.7617%	399	走	71.9619%	475	约	76.2164%	551	亲	79.6766%
324	球	66.8379%	400	指	72.0234%	476	星	76.2660%	552	爱	79.7181%
325	创	66.9139%	401	排	72.0846%	477	功	76.3156%	553	响	79.7596%
326	没	66.9896%	402	白	72.1457%	478	算	76.3650%	554	低	79.8010%
327	那	67.0652%	403	料	72.2066%	479	网	76.4144%	555	红	79.8424%
328	气	67.1405%	404	游	72.2676%	480	精	76.4636%	556	银	79.8837%
329	价	67.2154%	405	型	72.3283%	481	反	76.5128%	557	寄	79.9250%
330	知	67.2902%	406	做	72.3888%	482	春	76.5618%	558	培	79.9661%
331	格	67.3650%	407	许	72.4493%	483	护	76.6107%	559	族	80.0071%
332	样	67.4398%	408	施	72.5095%	484	村	76.6596%	560	益	80.0480%
333	级	67.5134%	409	儿	72.5697%	485	非	76.7083%	561	材	80.0889%
334	师	67.5868%	410	变	72.6296%	486	极	76.7568%	562	石	80.1297%
335	辑	67.6601%	411	副	72.6894%	487	列	76.8049%	563	克	80.1705%
336	给	67.7332%	412	奖	72.7491%	488	仅	76.8530%	564	采	80.2113%
337	视	67.8063%	413	义	72.8088%	489	病	76.9010%	565	祖	80.2519%
338	真	67.8793%	414	医	72.8683%	490	则	76.9489%	566	九	80.2925%
339	步	67.9516%	415	构	72.9276%	491	象	76.9968%	567	必	80.3328%
340	把	68.0237%	416	养	72.9867%	492	聘	77.0445%	568	友	80.3728%
341	想	68.0957%	417	协	73.0457%	493	域	77.0922%	569	吃	80.4127%
342	住	68.1671%	418	你	73.1044%	494	读	77.1397%	570	黄	80.4525%
343	认	68.2382%	419	快	73.1628%	495	势	77.1870%	571	销	80.4924%
344	优	68.3088%	420	林	73.2211%	496	越	77.2340%	572	千	80.5322%
345	百	68.3794%	421	线	73.2794%	497	织	77.2810%	573	未	80.5719%
346	李	68.4496%	422	完	73.3376%	498	况	77.3279%	574	归	80.6115%
347	标	68.5196%	423	志	73.3957%	499	称	77.3746%	575	买	80.6511%
348	需	68.5896%	424	届	73.4537%	500	双	77.4214%	576	命	80.6906%
349	供	68.6594%	425	告	73.5116%	501	马	77.4681%	577	存	80.7298%
350	号	68.7292%	426	营	73.5694%	502	何	77.5148%	578	按	80.7691%
351	打	68.7986%	427	积	73.6271%	503	众	77.5614%	579	随	80.8084%
352	造	68.8679%	428	它	73.6848%	504	即	77.6079%	580	图	80.8475%
353	影	68.9370%	429	色	73.7422%	505	容	77.6543%	581	故	80.8861%
354	类	69.0060%	430	版	73.7997%	506	超	77.7006%	582	座	80.9247%
355	战	69.0746%	431	直	73.8570%	507	源	77.7467%	583	思	80.9630%
356	形	69.1430%	432	准	73.9143%	508	德	77.7928%	584	落	81.0013%
357	感	69.2113%	433	栏	73.9714%	509	案	77.8385%	585	限	81.0393%
358	张	69.2796%	434	带	74.0286%	510	央	77.8842%	586	航	81.0773%
359	历	69.3478%	435	引	74.0856%	511	率	77.9299%	587	属	81.1151%
360	热	69.4158%	436	划	74.1427%	512	器	77.9755%	588	票	81.1528%
361	乐	69.4838%	437	史	74.1996%	513	声	78.0211%	589	牌	81.1903%
362	深	69.5517%	438	环	74.2565%	514	户	78.0666%	590	购	81.2278%
363	论	69.6195%	439	望	74.3132%	515	失	78.1120%	591	语	81.2653%
364	考	69.6874%	440	占	74.3694%	516	站	78.1570%	592	录	81.3027%
365	该	69.7551%	441	示	74.4256%	517	余	78.2020%	593	晚	81.3402%
366	具	69.8227%	442	续	74.4810%	518	显	78.2470%	594	额	81.3775%
367	纪	69.8903%	443	园	74.5363%	519	旅	78.2920%	595	满	81.4148%
368	质	69.9577%	444	速	74.5914%	520	远	78.3370%	596	份	81.4521%
369	博	70.0252%	445	兴	74.6465%	521	便	78.3818%	597	让	81.4893%
370	再	70.0921%	446	效	74.7013%	522	试	78.4267%	598	剧	81.5265%
371	居	70.1587%	447	始	74.7561%	523	负	78.4715%	599	律	81.5632%
372	亚	70.2250%	448	空	74.8107%	524	片	78.5160%	600	音	81.5999%
373	她	70.2912%	449	消	74.8654%	525	严	78.5604%	601	阳	81.6364%
374	见	70.3573%	450	权	74.9200%	526	确	78.6046%	602	欢	81.6728%
375	验	70.4234%	451	注	74.9739%	527	装	78.6488%	603	铁	81.7091%
376	己	70.4892%	452	食	75.0275%	528	闻	78.6929%	604	党	81.7453%
377	备	70.5551%	453	包	75.0804%	529	根	78.7370%	605	模	81.7815%
378	英	70.6208%	454	招	75.1333%	530	争	78.7810%	606	龙	81.8176%
379	身	70.6866%	455	较	75.1860%	531	段	78.8248%	607	庭	81.8536%
380	室	70.7520%	456	识	75.2385%	532	复	78.8686%	608	写	81.8895%

609	除	81.9254%	685	税	84.4872%	761	礼	86.6039%	837	拍	88.3622%
610	策	81.9612%	686	角	84.5177%	762	六	86.6292%	838	植	88.3832%
611	配	81.9970%	687	酒	84.5482%	763	街	86.6544%	839	幅	88.4043%
612	础	82.0328%	688	担	84.5786%	764	七	86.6796%	840	季	88.4253%
613	什	82.0685%	689	稿	84.6088%	765	免	86.7047%	841	丰	88.4462%
614	庆	82.1041%	690	灾	84.6389%	766	藏	86.7298%	842	树	88.4672%
615	厂	82.1397%	691	念	84.6688%	767	毕	86.7548%	843	粮	88.4881%
616	致	82.1753%	692	镇	84.6987%	768	左	86.7798%	844	沿	88.5089%
617	著	82.2107%	693	眼	84.7284%	769	朱	86.8048%	845	筑	88.5296%
618	岗	82.2462%	694	尽	84.7581%	770	泽	86.8297%	846	印	88.5503%
619	习	82.2815%	695	终	84.7877%	771	独	86.8545%	847	野	88.5710%
620	健	82.3169%	696	邮	84.8172%	772	罗	86.8792%	848	厅	88.5917%
621	湖	82.3523%	697	药	84.8468%	773	右	86.9038%	849	刚	88.6122%
622	维	82.3875%	698	减	84.8763%	774	脑	86.9283%	850	毒	88.6328%
623	范	82.4228%	699	楼	84.9057%	775	画	86.9529%	851	探	88.6532%
624	卫	82.4579%	700	括	84.9350%	776	须	86.9774%	852	杨	88.6735%
625	言	82.4929%	701	献	84.9640%	777	核	87.0018%	853	攻	88.6939%
626	土	82.5280%	702	良	84.9929%	778	沙	87.0262%	854	孙	88.7143%
627	油	82.5631%	703	测	85.0217%	779	货	87.0504%	855	移	88.7346%
628	贸	82.5981%	704	依	85.0505%	780	微	87.0746%	856	励	88.7549%
629	像	82.6330%	705	财	85.0793%	781	抗	87.0987%	857	击	88.7752%
630	轻	82.6680%	706	修	85.1080%	782	桥	87.1228%	858	遇	88.7954%
631	降	82.7029%	707	苏	85.1367%	783	临	87.1469%	859	陆	88.8157%
632	富	82.7377%	708	歌	85.1653%	784	予	87.1709%	860	简	88.8359%
633	菜	82.7726%	709	困	85.1940%	785	波	87.1949%	861	害	88.8561%
634	防	82.8074%	710	善	85.2226%	786	停	87.2189%	862	补	88.8763%
635	素	82.8421%	711	男	85.2511%	787	宣	87.2429%	863	巴	88.8965%
636	层	82.8769%	712	适	85.2796%	788	压	87.2667%	864	死	88.9166%
637	态	82.9116%	713	秀	85.3081%	789	坚	87.2903%	865	射	88.9366%
638	介	82.9462%	714	午	85.3366%	790	股	87.3138%	866	龄	88.9566%
639	离	82.9808%	715	奥	85.3649%	791	班	87.3373%	867	绝	88.9765%
640	却	83.0154%	716	币	85.3932%	792	私	87.3608%	868	虽	88.9965%
641	边	83.0499%	717	汇	85.4214%	793	绩	87.3842%	869	川	89.0164%
642	景	83.0843%	718	置	85.4496%	794	纷	87.4076%	870	婚	89.0363%
643	底	83.1187%	719	钟	85.4778%	795	促	87.4310%	871	含	89.0562%
644	且	83.1530%	720	练	85.5059%	796	卖	87.4543%	872	般	89.0760%
645	预	83.1872%	721	吸	85.5340%	797	岛	87.4776%	873	舞	89.0958%
646	售	83.2213%	722	送	85.5621%	798	逐	87.5008%	874	湾	89.1156%
647	破	83.2553%	723	课	85.5901%	799	签	87.5240%	875	松	89.1353%
648	照	83.2892%	724	派	85.6180%	800	挥	87.5471%	876	互	89.1549%
649	云	83.3230%	725	津	85.6459%	801	荣	87.5701%	877	乎	89.1745%
650	火	83.3568%	726	血	85.6738%	802	疗	87.5931%	878	顺	89.1939%
651	早	83.3905%	727	孩	85.7017%	803	执	87.6161%	879	损	89.2133%
652	群	83.4242%	728	订	85.7294%	804	登	87.6391%	880	监	89.2326%
653	承	83.4580%	729	例	85.7570%	805	检	87.6620%	881	询	89.2519%
654	址	83.4916%	730	迎	85.7846%	806	差	87.6850%	882	扩	89.2711%
655	访	83.5253%	731	状	85.8120%	807	章	87.7078%	883	苦	89.2900%
656	洋	83.5588%	732	冠	85.8395%	808	宝	87.7306%	884	络	89.3089%
657	断	83.5923%	733	汽	85.8667%	809	某	87.7532%	885	姓	89.3278%
658	拉	83.6258%	734	稳	85.8939%	810	卡	87.7759%	886	蒙	89.3466%
659	险	83.6591%	735	康	85.9210%	811	曲	87.7982%	887	阶	89.3654%
660	福	83.6925%	736	训	85.9481%	812	杂	87.8206%	888	异	89.3841%
661	胜	83.7258%	737	令	85.9751%	813	愿	87.8429%	889	雨	89.4028%
662	讲	83.7591%	738	驻	86.0020%	814	贵	87.8652%	890	夜	89.4215%
663	待	83.7922%	739	讨	86.0289%	815	觉	87.8874%	891	巨	89.4402%
664	继	83.8254%	740	止	86.0557%	816	吴	87.9095%	892	鱼	89.4588%
665	店	83.8585%	741	控	86.0823%	817	拿	87.9316%	893	杯	89.4772%
666	陈	83.8915%	742	夫	86.1089%	818	短	87.9537%	894	束	89.4957%
667	汉	83.9242%	743	融	86.1355%	819	田	87.9756%	895	择	89.5140%
668	突	83.9565%	744	刘	86.1619%	820	换	87.9976%	896	幕	89.5323%
669	官	83.9887%	745	察	86.1884%	821	假	88.0195%	897	措	89.5506%
670	普	84.0207%	746	仍	86.2148%	822	纳	88.0413%	898	析	89.5689%
671	曾	84.0525%	747	候	86.2413%	823	征	88.0631%	899	污	89.5871%
672	劳	84.0840%	748	邓	86.2677%	824	丽	88.0848%	900	堂	89.6052%
673	兰	84.1154%	749	紧	86.2939%	825	找	88.1064%	901	拥	89.6233%
674	希	84.1468%	750	顾	86.3201%	826	启	88.1278%	902	船	89.6414%
675	您	84.1782%	751	另	86.3463%	827	饭	88.1493%	903	祝	89.6595%
676	审	84.2094%	752	味	86.3723%	828	母	88.1707%	904	威	89.6776%
677	围	84.2407%	753	充	86.3983%	829	黑	88.1921%	905	餐	89.6956%
678	升	84.2718%	754	温	86.4243%	830	套	88.2135%	906	永	89.7135%
679	飞	84.3029%	755	听	86.4502%	831	密	88.2348%	907	退	89.7315%
680	切	84.3338%	756	洪	86.4761%	832	钢	88.2561%	908	繁	89.7494%
681	宁	84.3648%	757	久	86.5018%	833	彩	88.2774%	909	典	89.7672%
682	绍	84.3955%	758	喜	86.5274%	834	竞	88.2987%	910	木	89.7849%
683	洲	84.4261%	759	激	86.5529%	835	急	88.3199%	911	召	89.8026%
684	武	84.4567%	760	细	86.5784%	836	烈	88.3411%	912	托	89.8203%

913	朋	89.8380%	930	智	90.1364%	947	烟	90.4273%	964	叶	90.7096%
914	鼓	89.8557%	931	痛	90.1537%	948	帮	90.4442%	965	欧	90.7260%
915	窗	89.8734%	932	诉	90.1710%	949	牛	90.4610%	966	振	90.7424%
916	遗	89.8910%	933	澳	90.1883%	950	佳	90.4777%	967	露	90.7588%
917	染	89.9087%	934	呢	90.2056%	951	序	90.4945%	968	惠	90.7751%
918	鲜	89.9263%	935	奇	90.2228%	952	享	90.5111%	969	播	90.7915%
919	伟	89.9440%	936	笔	90.2400%	953	软	90.5277%	970	韩	90.8078%
920	筹	89.9616%	937	犯	90.2572%	954	码	90.5443%	971	峡	90.8241%
921	峰	89.9792%	938	坐	90.2743%	955	跨	90.5609%	972	伤	90.8403%
922	吨	89.9968%	939	谢	90.2915%	956	绿	90.5775%	973	盛	90.8565%
923	草	90.0144%	940	夺	90.3086%	957	够	90.5941%	974	努	90.8726%
924	刻	90.0320%	941	毛	90.3257%	958	圳	90.6107%	975	冷	90.8887%
925	辆	90.0495%	942	救	90.3428%	959	怎	90.6272%	976	肉	90.9049%
926	轮	90.0670%	943	阿	90.3598%	960	似	90.6437%	.	.	. . . . .
927	夏	90.0844%	944	缺	90.3767%	961	库	90.6602%	4804	鼬	99.9888%
928	输	90.1017%	945	判	90.3936%	962	胞	90.6767%	.	.	. . . . .
929	略	90.1191%	946	旧	90.4105%	963	父	90.6931%	5229	膘	100.0000%

A szóstatisztikákhoz megbízható méretű korpusz kell. Szemben a karakterstatisztikákkal, itt nagyobb eltérés mutatkozik szerzők szerint külön készített elemzésekkel. Ez attól van, hogy mindenkinek meg van a saját nyelve, amiben a valószínűbben előforduló elemek gyakoriságai is lényegesen különbözhetnek. Elemezni persze azt tudjuk, ami le van kódolva, tehát csak az aktív nyelvet. A passzívat – tehát azt, mit mennyire értünk – adatok híján nem ismerjük. Mivel a ritkán (egyszer-kétszer) előforduló elemek statisztikája nem megbízható, a vizsgált szavak többségének tapasztalati gyakorisága szinte semmit nem mond a nyelvben elfoglalt valószínűségükről.

A szóstatisztika – mivel nagy számosságú halmaz elemei gyakoriságát vizsgáljuk – alkalmas szövegek szerzőinek azonosítására. Ha itt akarunk szó n-eseket felmérni, akkor már sokkal több különböző elemet kell figyelembe venni.

A szó n-esek statisztikájához már végképp reménytelen közvetlen mátrixot használni. A szópárok előfordulása is – elvi számuk nagysága miatt – nagyon ritka mátrixot generálna. Itt ráadásul az is gond lehet, hogy szavakkal, és nem egyszerű kódokkal kell indexelni a mátrixot.

Feladat 1: Készítsen szóstatisztikai programot, vagyis olyat, amely megszámlálja, mely szóalak hányszor fordul elő egy szövegben. A szavakat rakja gyakorisági sorrendbe, és vizsgálja, hány szó fedi le a szöveg felét, háromnegyedét, hétnyolcadát…

Feladat 2: Nézze meg, mennyiben különbözik különböző írók szóstatisztikája. Hasonlóan vesse össze egy író különböző műveinek szóstatisztikáját.

Feladat 3: Meg lehet-e becsülni egy nyelv szavainak számát a szóstatisztika alapján?

4.5 Flektáló és ragozó nyelvek

A nyelvek minőségi különbséget mutatnak abban a tekintetben, hogyan változtatják alakjukat. Egyes nyelvekben a szó csak egyetlen vagy nagyon kevés formában jelenhet meg. Ilyen például a kínai, de az angolt is ide sorolhatjuk. Ezek a nem, vagy alig ragozó nyelvek.

Ezzel szemben a flektáló nyelvekben a szavaknak jellegzetesen több alakjuk van. A sémi nyelveknél a szótőt megelőzheti egy-két előtag (prefix), követheti egy-két toldalék (suffix), és mindeközben változik magának a tőnek a formája. Alapvetően a tövön belüli magánhangzók módosulnak. Ezeket a módosulásokat felfoghatjuk mint infix.

Gyakoribb, amikor a szótő egyáltalán nem vagy nem lényegesen változik, hanem csak toldalékok tapadnak a szavakra. Ezek a ragozó nyelvek. A ragozó nyelvek többségénél a toldalék a szótőt követi, de egy szóhoz egyszerre csak meghatározott (egy-két) toldalék kapcsolódhat, meghatározott sorrendben. Az ilyen nyelveket (pl. görög, olasz, orosz) ragozó nyelveknek hívják.

Ennél összetettebb toldalékolási struktúrával az agglutinatív nyelvek rendelkeznek, ahol is a toldalékok, bár megadott szabály szerint, de több rétegben tapadnak a szótőhöz. Ilyen a magyar, de a finn és a török is ehhez a csoporthoz tartozik. Ebből ered, a nagyszámú szóalak.

A két véglet (a nem ragozó és az agglutinatív) között az átmenet folyamatos. A határt meghúzni ott kell, ahol a praktikum megköveteli, vagyis az, ahogy a szavak kezelése technikailag megoldható. Az angol is felfogható ragozó nyelvnek, és a német nyelvben is van olyan tőváltozás (umlaut), amelyik nem a morféma határa környékén megy végbe. Ennek ellenére a ragozás nem alapvető ezeken a nyelveken.

A nem vagy gyengén ragozó nyelvek mondatszerkezetében nagyobb szerepet játszik a szavak sorrendje, hiszen a szavak, mondatrészek közötti kapcsolat milyenségére nincs más eszköz, mint a szavak szigorú

sorrendje. Emiatt ezek a nyelvek kötött szórendűek. Ezzel szemben az erősen ragozó nyelvekben a szórend szabadabb, hiszen az összefüggésekre a szavak ragozott alakjai is utalnak.

4.6 Szóelemzés asszociációs lista alapján

A morfológiai elemzést a hagyományosan úgy képzelték el, hogy minden szóalakhoz hozzá kell csak rendelni, melyik lexémát milyen nyelvi tulajdonsággal ruháztak fel, hogy megkapja a kívánt alakot (elemzés) vagy megfordítva, egy lexéma adott nyelvi tulajdonságokkal hogyan jelenik meg a felszínen (generálás). Sok nyelvi munka nem is foglalkozik sokat a morfológiával, feltételezik, hogy évtizedek alatt a nyelvészek feldolgozták a nyelv szóalakjait, és megadták a szóalakok lehetséges elemzését. Ezt a módszert természetes módon lehet alkalmazni az angol nyelv esetén:

flies = *fly* (*légy* főnév többes szám) | *fly* (*repül* ige egyes szám 3. személy)

Német nyelven: *Gift* = *der* ~ (*ajándék*) | *die* ~ (*düh*, *méreg*)

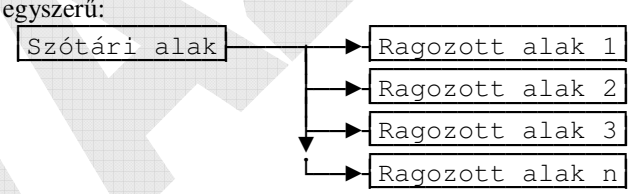
Magyarul, ha lenne ilyen:

lépnek = *lép*(ige, kijelentő mód, alanyi ragozás, többes szám 3. személy) | *lép*(főnév, részeseset)…

A mintáimban tudatosan vettem olyan szavakat, melyeknek több értelmezése van. Sem a szóalak nem határozza meg a szó elemzését egyértelműen, sem egy szó nyelvtani tulajdonságokkal való felruházása nem határozza meg a szó alakját, bár ebben az irányban ritkább a választási lehetőség (*alapul-alapszik*, *kettő-két*, *azzal-avval*, *mondta-mondotta*…). A fenti példák arra szolgálnak, hogy megmutassam, minden szóalakhoz hozzá kell rendelni a szó szótári alakját és azokat a nyelvtani tulajdonságokat, amelyek fontosak lehetnek a további céloknak, pl. mondatelemzéshez. A kérdés az, hogy ezt a hozzárendelést hogyan valósíthatom meg. Ha a szóalakok száma nem vészesen nagy, akkor minden szóalak egy-egy kulcsszónak fogható fel, és az informatikában szokásos indexelési technikák tökéletesen megfelelnek, hogy a szóalakhoz hozzárendeljük a külön tárban levő elemzéseket. Ez az elemző leírás, Általában ezt fordítva generatív módon szokták leírni, vagyis a szótári alakhoz rendelik a szó egyéb formáját, de a hozzárendelés könnyen megfordítható. Egy angol helyesírási szótár a következőképpen teszi (a szón belüli függőleges vonás a szótagolás helyét jelöli):

wolf, wolves, wolf's, wolves'
wo|man, wo|men, wo|man's, wo|men's
work, works, worked, work|ing
work, works, work's, works'
works, works' (pl.)
work|er, work|ers, work|er's, work|ers'

A modell nagyon egyszerű:



A módszer akkor hatékony, ha egy szó alakjainak száma a kognitív korlát alatt marad. Ez az angolnál így van. Ha nem így lenne, az adatbázis karbantartása lehetetlenné válna. Amennyiben csak szóellenőrzés a cél, akkor elegendő a szóformákat egy alkalmas tömör, de gyors kereséssel elérhető struktúrában tárolni. Erre alkalmasak a ciklusmentes gráfok (DAG), de más módszerek is hatékonyak lehetnek.

Ha a szavak elemzése is fontos, akkor minden szóformához egyedi (esetleg alternatív) nyelvi információt kell rendelni. Emiatt a ciklusmentes gráf nem alkalmas a tároláshoz. A szófa alkalmazása már lehetővé teszi az egyedi információk hozzárendelését a szóalakhoz, de ehhez a fa – mint gráf – csúcsinak száma legalább annyi, mint ahány szóalak van a nyelvben. Angolszász nyelveknél ennek ellenére megfelelő megoldás, ha a szóalakokhoz (a szófa csúcsaihoz) hozzárendeljük a szó elemzését, mivel egy szóalaknak nem szokott túl sok alternatív elemzése lenni. Az elemzéseket is karakterláncként reprezentálhatjuk. Így az elemzési karakterláncokat is tárolhatjuk szófában. Akár elemzésről, akár generálásról van szó, csupán azt kell megoldani, hogy egy fa gyökerétől egy pontjáig vezető úthoz hozzárendeljünk egy másik fa egy vagy több útját.

Tegyük fel, hogy ilyen módszerrel akarunk szótárat építeni. Ha másodpercenként 1 szót kódolunk, és 1000 000 szó van, valamint éjjel-nappal dolgozunk, akkor két hét alatt kész vagyunk. Persze, ha a hibaszázalék a fáradtsággal nő, akkor kódolásunk eredménye már az első óra után siralmas lesz, Ha ezt nem

feltételezzük, mert kódolásban tökéletesek vagyunk, meddig tart, ha a szóalakok száma több mint 1000 000 000!?

Egyébként is, emberek által készített nyelvi adatbázis egy méret felett meghatározható hibaszázalékot tartalmaznak. Ha igaz, hogy az adatbázis elemei közül azok, melyek használata gyakoribb, azok nagy valószínűséggel helyesen kerülnek be az adatbázisba, akkor a ritkáknak hibás felvétele nagyobb eséllyel történik. Sajnos a hibák javítására is akkor van nagyobb esély, ha azokat gyakrabban előforduló elemeknél vétjük el. Sőt, ha létezik a hibák feltárására valamilyen taktika, abban is lehetnek hibák. Ennek következménye, hogy egy méret után gyakorlatilag nem fejleszthető egy nyelvi adatbázis. Csak akkor, ha a korrigálási/bővítési módszeren tudunk változtatni, vagy az adatbázist alakítjuk át úgy, hogy kevesebb adattal többet tudunk kódolni. A tapasztalat szerint egy szakember legfeljebb 10 000 tételes adathalmaznál éri el azt a határt, ami felett a hibaszázalék csak nőhet. Ha ezt ügyesen szervezett csoportmunkában végzik, tehát valamilyen keresztellenőrzésre is van lehetőség, akkor 200 000, 250 000 tétel a maximum Efelett már a hibaszázalék nem javul, inkább romlik.

4.7 Szóelemzés konkatenációs morfológia esetén

Az előbbiek alapján, ha az egyes szavak alakjainak száma tíznél több, vagyis a szóalakok száma 100 000-nél nagyobb, akkor a szavakat mindenképpen érdemes felbontani funkcionális részekre, már ha a nyelv szótana ezt lehetővé teszi. Ha a nyelv szavai jól darabolhatók részekre, akkor fel kell tárni, milyen elemek melyeket követhetnek. Így a szótár nem a szavak alakjait, hanem az alkotórészek, a morfémák alakjait kell tárolnia. Ez nagyságrendekkel csökkenti a szótár(ak) méreteit. Német nyelv esetén, ha az összes lehetséges szóösszetételt tárolnánk, akkor nem tíz-húszezer szót kéne tárolni, hanem milliárdokat. Magyar nyelv esetén is gondot okozna az összetett szavak számbavétele, de a nélkül is bajban vagyunk a toldalékolás összetettsége miatt. Szerencsére a toldalékok jól leírható sorrendben követik egymást, így szabályokkal írhatjuk le, mi után mi következhet.

A részek egymáshoz kapcsolódása még inkább követhető a latin és a szláv nyelveknél. Ha egy orosz, lengyel vagy spanyol szótárat veszünk a kezünkbe, akkor a szótári tételek, a szavak mellett egy olyan kódszámot találhatunk, amely egyértelműen megadja, hogy a szó milyen toldalékokkal látható el. Egy ilyen osztályhoz tartozó toldalékcsoportot felfoghatunk egy kis segédszótárnak. Ezek után csak a szótövet és a folytatási osztályban levő toldalékot kell összefűzni.

Összefűzni nem a szótári alakot kell a toldalékkal, hanem általában az ettől eltérő szótövet. Orosz, német szótárakban az ige szótári alakja, az infinitívusz forma. Ez nem csak a tövet, hanem egy toldalékot is magába foglaló alak. Ezt levágva kapjuk meg a szótövet, és a szótőhöz kell hozzáilleszteni a lehetséges toldalékot.

Tovább bonyolítja a dolgot, hogy a szónak nem feltétlen egy tőalakja van. sőt a különböző tőalakok más-más toldalékokkal folytatódhatnak. Ez azonban a probléma összetettségét kevésbé növeli ahhoz képest, mintha az összes ragozott alakot kéne egyenként felsorolni. A modell a következő sémával írható le:

szótári alak+ragozási osztály → szótőalak+folytatási osztály

Vegyünk egy egyszerűsített példát, és próbáljuk meg az angol nyelv főnévi ragozásának egy részét így megadni:

ragozási osztály	lemma	folytatási osztály	form	singular, no prop.	singular, property	plural, no prop.	plural, property
N1	*boot*	N1	*boot*		*'s*	*s*	*s'*
N2	*fox*	N2	*fox*		*'s*	*es*	*es'*
N3	*fly*	N3a	*fly*		*'s*	-	-
		N3b	*fli*	-	-	*es*	*es'*
N4	*matrix*	N4a	*matrix*		*'s*	*es*	*es'*
		N4b	*matrice*	-	-	*s*	*'s*
N5	*mouse*	N5a	*mouse*		*'s*	-	-
		N5b	*mice*	-	-		*'s*
N6	*wife*	N6a	*wife*		*'s*	-	-
		N6b	*wive*	-	-	*s*	*s'*
N7	*ox*	N6	*ox*		*'s*	*en*	*en's*
N8	*datum*	N8a	*datum*		*'s*	-	-
		N8b	*data*	-	-		*'s*

N9	*medium*	N8a	*medium*		*'s*	*s*	*s'*
		N8b	*media*	-	-		*'s*

A táblázat könnyen folytatható, igei, melléknévi osztályok is lekódolhatók. Ilyen modell megalkotásánál nem kell az összes alakját felsorolni a szavaknak, viszont meg kell adni, hogy a szó szótári alakjából hogyan származtatható a szótő (esetleg szótőalakok) és az adott szótőalakhoz milyen toldalékok kapcsolódhatnak. Ez az angolnál nem túl nehéz feladat, de a magyar nyelv esetén már nyelvésznek is komoly munkát okoz. (Lásd a 4.10 fejezetet.)

Előnye, hogy ilyen leírásból aránylag könnyen készíthető szóellenőrző, -elemző vagy -generátor modul. Sokkal tömörebb, logikusabb, mint a listás módszer. Hátránya, hogy azoknál a nyelveknél, amelyeknél a szó toldalékai nem egy, hanem mindkét oldalról kapcsolódhatnak, esetleg infix – belső változás is előfordul a szószerkezetben, nehezen, vagy egyáltalán nem alkalmazható.

4.8 Egy egyszerűsített elemző modell

Ha a morfológiai leírás alapvetően elemzési céllal készül, akkor egy gyors heurisztikus módszer is célravezető lehet.

Az alapvető igény onnan származik, hogy – bár a nyelvtani, toldalékolási szabályok, a toldalékok halmaza, azok nyelvtani szerepe könnyen leírható, a szókészlet mérete miatt sohasem lehet teljes a szavak osztályozása. A szókészlet leírása, annak nyelvi kódolása gondot jelenthet. Ha viszont elegendő a szavak szótöveit (esetleg szótári alakjait) összegyűjteni, és ezekhez csak annyi ismeretet tenni, hogy mi a szófajuk, akkor a szótár építése, karbantartása a felhasználóra bízható. Hogyan lehet ebből elemzőt, helyesírás-ellenőrzőt készíteni?

Nézzünk egy példát az angolra:

1. Ha egy szóalak szerepel a szótárban, akkor az alapszó.
2. Ha *'s*-re végződik, akkor vágjuk le az *'s*-et, és nézzük meg, szerepel-e a maradék a szótárban. Ha igen, akkor (egy-két kivételtől eltekintve) ez egy főnév, ami birtokjellel rendelkezik.
3. Ha nem *'s*-re, csak *s* re végződik, akkor ennek levágásával újra vizsgáljuk meg a maradékot. Ha ez szerepel a szótárban, akkor aszerint, hogy a szótári tétel igető vagy főnévi tő, a szó egyes szám harmadik személyű vagy többes számú alakját találtuk meg.
4. *-es* levágása is hasonló működik, mint az előző.
5. *-es* levágása esetén, ha *i*-re végződik a maradék, írjuk át az *i*-t *y*-ra, és úgy ellenőrizzük a szót a szótárban.

Stb.

Ezt a durva modellt helyesírás-ellenőrzőként nem ajánlatos alkalmazni, mivel túlgenerál, vagyis elfogad nem létező alakokat is. Ezzel szemben, ha célunk az, hogy jól írt szövegeket elemezzünk, az ebből eredő félreelemzések száma elhanyagolható (az angol nyelv esetén). Helyigénye kisebb, mint az előző fejezetben használtnál, könnyen karbantartható, nincs szükség nyelvészre a szótár bővítésénél. Természetesen a rendhagyó ragozású szavak minden alakját fel kell venni a szótárba, de mivel ezek száma kicsi, nem okozhat gondot.

Ha magyar nyelvre akarunk hasonló módszert, akkor az nem ennyire egyszerű. A cél, hogy olyan leírást adjunk, amelyben a szótár bővítésére elegendő a szavak szótári alakját felvenni, esetleg szófaji beosztásukkal együtt. Lásd a magyar szavak elemzésénél: 4.12.1.

4.9 Ispell, Myspell és hasonló módszerek

A konkatenációs modell jól algoritmizálható. Egy egyszerű változata a szabad szoftver világban használatos Ispell és rokonai (Myspell, Hunspell). Ebben jól szeparálódik a szótár és a toldaléktár. Mindkettőt egyszerű szöveglistában adják meg, és ez a mód alkalmas arra, hogy sokan sokszor javítgassák, bővítsék. A szótárban a szavak mellett jelezni lehet, hogy milyen toldalékcsoportot kapcsolhatunk a szóhoz. Akár többet is. A toldaléktárban a csoportok definíciója mellett egyszerű reguláris kifejezésekkel lehet megadni, milyen esetben, milyen tőváltozást kell végrehajtani a toldalék kapcsolódási felületénél. Két szó egy spanyol szótárból:

potenciar/REÄDŔ acanto/aS

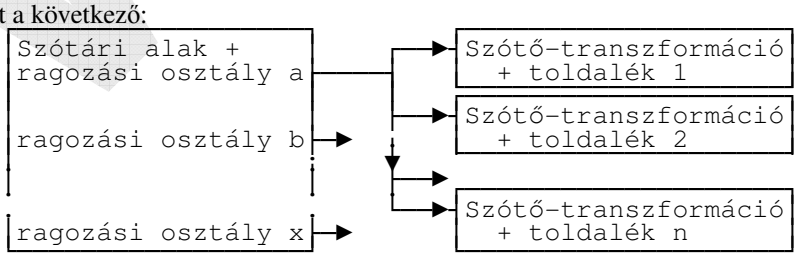
A szavakat szótári alakban kell megadni, és utána jól elkülönítve egy-egy karakterrel kell megnevezni azokat az osztályokat, melyek a toldaléktárban fejtenek ki.

Néhány tétel a toldalékleírásból:

```
PFX a Y 2
PFX a 0 a [^aeiou]
PFX a 0 an [aeiou]

SFX R Y 197
SFX R r mos [aei]r
SFX R ar áis ar
SFX R r ba ar
SFX R r bas ar
SFX R ar ábamos ar
SFX R ar abais ar
SFX R r ban ar
SFX R ar é [^cguz]ar
SFX R car qué car
SFX R ar ué gar
SFX R ar é [^g]uar
SFX R uar üé guar
SFX R zar cé zar
SFX R r ste ar
SFX R ar ó ar
SFX R r steis ar
SFX R r ron ar
SFX R 0 é [aei]r

SFX E Y 73
SFX E ar o ar
SFX E r s [ae]r
SFX E r 0 [ae]r
SFX E r n [ae]r
SFX E ar e [^cguz]ar
SFX E car que car
SFX E ar ue gar
SFX E ar e [^g]uar
SFX E uar üe guar
SFX E zar ce zar
SFX E ar es [^cguz]ar
SFX E car ques car
SFX E gar gues gar
SFX E ar es [^g]uar
SFX E uar ües guar
SFX E zar ces zar
SFX E ar en [^cguz]ar

SFX Ä Y 16
SFX Ä ar ándola ar
SFX Ä ar ándolo ar
SFX Ä ar ándolas ar
SFX Ä ar ándolos ar
SFX Ä er iéndola [^aeo]er
SFX Ä er iéndolo [^aeo]er
SFX Ä er iéndolas [^aeo]er
SFX Ä er iéndolos [^aeo]er
SFX Ä er yéndola [aeo]er
SFX Ä er yéndolo [aeo]er
SFX Ä er yéndolas [aeo]er
SFX Ä er yéndolos [aeo]er
SFX Ä r éndola ir
SFX Ä r éndolo ir
SFX Ä r éndolas ir
SFX Ä r éndolos ir

SFX D Y 12
SFX D r do [aií]r
SFX D r dos [aií]r
SFX D r da [aií]r
SFX D r das [aií]r
SFX D er ido [^aeo]er
SFX D er ído [aeo]er
SFX D er idos [^aeo]er
SFX D er ídos [aeo]er
SFX D er ida [^aeo]er
SFX D er ída [aeo]er
SFX D er idas [^aeo]er
SFX D er ídas [aeo]er

SFX Ŕ Y 4
SFX Ŕ 0 la r
SFX Ŕ 0 lo r
SFX Ŕ 0 las r
SFX Ŕ 0 los r

SFX S Y 32
SFX S 0 s [aceéfgiíkoóptuúw]
SFX S 0 es [bdhíjlmrúy]
SFX S á aes á
SFX S 0 es [^áeéíóú]n
SFX S 0 es [^áéíóú]s
SFX S án anes án
SFX S én enes én
SFX S ín ines ín
SFX S ón ones ón
SFX S ún unes ún
SFX S ás ases ás
SFX S és eses és
SFX S ís ises ís
SFX S ós oses ós
SFX S ús uses ús
SFX S 0 es [^dgmv]en
SFX S orden órdenes orden
SFX S agen ágenes agen
SFX S argen árgenes argen
SFX S igen ígenes igen
SFX S irgen írgenes irgen
SFX S amen ámenes amen
SFX S armen ármenes armen
SFX S emen émenes emen
SFX S ermen érmenes ermen
SFX S imen ímenes imen
```

A toldaléktárban PFX, illetve SFX kulcsszó jelzi, hogy előtag vagy utótag szerepel a tételben, majd a toldalékosztály egykarakteres neve, a toldalék(forma), ezt követi az illeszkedési feltétel egyszerű (korlátozott) reguláris kifejezéssel, végül, ha kell, az esetleges tőváltozást és toldalékformát megadó helyettesítési szabály. Ez a fajta leírás a felhasználóknak kevésbé szemléletes, nehézkes, és módosítása gyakran vezethet újabb hibához. A felhasználók csak a szótárat bővítik, a nyelvtani (toldalékolási) részhez nem nagyon nyúlnak. A közepesen toldalékoló nyelveknél azért ez a formalizmus is kellőképpen karbantartható szakavatott által. Bonyolultabb morfológiánál már mindenképpen szükséges, hogy ne ezt a leírást változtassuk közvetlenül, hanem egy más, áttekinthetőbb leírást használjanak, és ebből generálják a fenti formalizmusnak megfelelő kódot. Korlát még, hogy az olyan tőváltozást, ami nem közvetlen a toldalék csatlakozásánál történik – pl. a német főneveknél az *a-ä, u-ü* átalakulás, nem írható le jól. A többszörös toldalékolások sem kezelhetőek, ezért magyarra alkalmazni nem a legszerencsésebb formalizmus. A szóösszetételre sem tökéletes a leírás. A szószablya projektben ezeket a korlátokat sikerült valamennyire legyőzni.

A modell itt a következő:



4.10 A magyar nyelv morfológiája

A magyar nyelv tipikusan agglutináló nyelv. Ez azt jelenti, hogy a szavakhoz több rétegben rakódnak végződések. Nyelvünkben néhány fajta előtag is szerepet játszik. Ezek az igekötők, melyek egyes esetekben különválhatnak a szótól, és a melléknevek felső és túlzófokát kifejező *leg-* illetve *legesleg-* előtag. Az előtagok alkalmazása több szempontból olyan, mint a szóösszetétel - pl. szóelválasztás szempontjából - míg a toldalékok helyes használata sok gondot okoz úgy a magyar anyanyelvűeknek, mint a magyarul megtanulni szándékozóknak.

A szóalaktani modell feladata, hogy a szavakat funkcionális részekre vágja, feltárja szerkezetét és módot adjon számítástechnikai szempontból is kezelhetővé tenni mind elemzés mind generálás céljából. A magyar nyelvnél gyakran alkalmazott módszer, hogy a szavakat jelentéssel bíró morfémákra, és jelentés nélküli fonetikai töltelékre (kötőhangokra) bontják. Kérdés persze, hogy mik a morfémák, és mik a kötőhangok. Én ezt a kérdést megkerülöm, és a toldalékok részének tekintem az összes hozzátartozó fonémát. Ily módon egy morfémának számos megjelenési formája van. Azzal a kérdéssel, hogy egy végződés mikor milyen alakot ölt a felszínen, a morfológiai elemzés kapcsán részletesebben foglalkozom, de hogy miért, arról nem kívánok elméleti meggondolásokba mélyedni. Jelen fejezetben a toldalékok funkcionális tulajdonságairól szeretnék képet rajzolni.

A toldalékok hagyományos felosztása: ragok, jelek, képzők. A szokásos (jelentéstani) magyarázat szerint: a képzők megváltoztatják a szó jelentését, a jelek módosítják azt, míg a ragok a szónak a mondatbeli viszonyát határozzák meg.

Antal László "Egy új magyar nyelvtan felé" című művében ennél pontosabb kritériumot ad a felosztáshoz. A feltételek alapvetően formaiak, és másodsorban szemantikaiak. Névszók esetén szerinte:

1. A jelek egyértelműen felsorolhatók: a többes szám jele, a birtokragok és a birtokos jel.
2. A rag olyan kötött toldalék, mely megjelenhet mindenféle jel (és képző) után, nincsenek tiltott szemantikai sávjai, és a ragot nem követheti semmilyen más toldalék.
3. A képzők kötött toldalékok, melyek társulhatnak más képzőkkel, nem léphetnek fel jelek után, és vannak szemantikai tiltó sávjaik.

Az én kiindulási alapom hasonló, bár alapvetően szintaktikai alapú. Egy toldalékot aszerint sorolok be, hogy:

– hol a helye a szóban,
– milyen szavakra alkalmazható,
– milyen szerepe van a mondatban

4.10.1 A szófaj mint toldalékolási kategória

A szavak nyelvi szempontból szófajokba soroltatnak. Ennek a besorolásnak két szempontja van. Az egyik a szótani, a másik a mondattani. Angolul a szófajt *part of speach*-nek nevezik, ami a második szempontot hangsúlyozza. Náluk a szóalaktan kisebb szerepet játszik. A mondattani szempontokkal itt kevesebbet foglalkozunk, de időnként utalok arra is. Most fontosabb, mely szóhoz milyen toldalékok csatlakozhatnak.

A hagyományos beosztás szerint van *ige*, *főnév*, *melléknév*, *számnév*, *névmás*, és egyebek (*kötőszavak*, *névutók*, *határozók*…). A beosztás nem túl szerencsés, mert a névmások nem önálló csoport, hanem egyes elemei a főnévekhez, mások melléknevekhez, míg van, amelyik a számnevekhez sorolandók, és ez sokkal jelentősebb, mint az, hogy névmás. Külön kezelni legfeljebb azért érdemes a szótan szempontjából, mert toldalékolásuk korlátozott. Másrészt a névszók toldalékolása egységes részt tartalmaz, ezért egy csoportba tehetjük, ha csak durva osztályokra van szükség, míg akár sok alosztályra bonthatjuk a főneveket, melléknéveket, számneveket, ha pontosítjuk a modellünket.

A ragozás szempontjából négy osztályt érdemes megkülönböztetni: névszók, igék, gyengén toldalékolódók és ragozhatatlanok. A pontosabb toldalékoláshoz – főként a képzők szempontjából – a klasszikus beosztás sem elég. Az ennél pontosabb felosztásnak legtöbb esetben mondattani szerepe is van.

4.10.2 Névszói (eset)ragok

A ragok közös tulajdonsága, hogy mindig a szó utolsó toldalékai, tehát a később említendő kivételeket nem tekintve a ragot már semmilyen toldalék nem követheti.

Az esetrag a névszó végén álló toldalék. Alkalmazható a névszavakra, névszóvá képzett szavakra egyaránt. A névszóból határozót formál, mely alapján a mondatban, vagy kisebb szintagmában határozói szerepet tölthet be a kérdéses szó, kifejezés (ideértve a tárgyat és alanyt is). Van olyan nyelv, amelyben csak a főnevek kaphatnak esetragot, a melléknevek nem. Magyarban a ragok többségét minden névszó megkaphatja.

Más nyelvekben – pl. a grúzban – az alanyesetű szó is ragot kap, így a szótő különbözik az alanyesetű formájától a szónak, és az önmagában nem is szerepel szövegben. Hasonló a jelenség, mint magyarban az ikes igéknél.

Általában az esetragok határozzák meg egy egész névszói csoportnak a szerepét a mondatban, vagy egy részkifejezésben. (ld. **5.3.1.** vonzatok, névszói kifejezések).

A különböző nyelvtani munkákban különböző számú esetragot tartanak nyilván. Én a következő 33-at tartom számon:

:	*piros,*	ig:	*nyolcig*,
t:	*pirosat*,	szor:*	*nyolcszor*,
ban:	*pirosban*,	szorta:*	*nyolcszorta*
ba:	*pirosba*,	szorra:*	*nyolcadszorra*
ból:	*pirosból*,	val:	*pirossal*,
an:*	*pirosan*,	nak:	*pirosnak*,
on:	*piroson*,	ul:*	*magyarul*,
ra:	*pirosra*,	stul:*	*pirosostul*,
ról:	*pirosról*,	képp(en):*	*eredményképpen*,
vá:	*pirossá*,	lag:*	*színleg*,
nál:	*pirosnál*,	onta:*	*naponta*,
hoz:	*piroshoz*,	tt:*	*Pécsett*,
tól:	*irostól*,	szerte:*	*országszerte*
ként:*	*pirosként*,	szám(ra):*	*zsákszám*,
ért:	*pirosért*,	rét:*	*nyolcrét*,
nként:*	*pirosanként*,	mód(on/ra):*	*gombamódra*
kor:*	*nyolckor*,		

Ez a lista több szempontból is eltér Antal Lászlóétól. Mint a fentiekből is látszik, ide soroltam mindazokat a végződéseket is, amit Antal László határozószói képzőnek tart (a *-gal jelölteket), olyan alapon, hogy nem alkalmazhatóak minden szóra, vagy nem állhat előttük jel. Ebből a szempontból ne azt tartsuk mérvadónak, hogy jelek megelőzhetik-e, hanem azt, hogy más toldalék már nem követheti, és határozót képeznek a szóból.

A számítógépes elemzés, során egyébként sok olyan furcsasággal találkoztam, ami nem nekem tűnt fel, hanem a gépi feldolgozás során derült ki. Az *-ul/-ül* rag például számos esetben állt birtokjel után (pl.: *fiául fogadta, társául szegődött*, stb.), sőt számos olyan ige van, aminek *-ul/-ül* vonzata van (pl.: *fogad, választ, beszél* stb.). Tehát ezek szerint rag. De találkoztam olyan mondattal is, hogy:

Pontjainkként egy-egy merőlegest állítva…

Az sem látszik teljesen igaznak, hogy a tisztán ragnak tekintettek bármikor alkalmazhatóak szemantikai ismérvek nélkül (pl.: ?*eléghez*, **tegnapban*, ***otthonon*, holott *eleget*, *tegnapra*, *otthonról* alak van). Inkább az lehet, hogy a ragok használata is esetleges, bár kevésbé, mint a képzőké. Időnként még határozószók, igék is kapnak esetragot, de ezt kivételnek tekintjük. Viszont például a *-szor*, *-rét* szinte kizárólag számnévhez fűzhető, a *-szám* csak mértékegységgel (esetleg számmal) használható, az *-an* pedig többnyire melléknévvel. Bár egyes példában csak számnévnek tulajdonítható ragot egyes főnevek is felvehetnek:

A terület egyenlő hosszor szélesség

Ebből következően az esetragokat tovább lehetne csoportosítani aszerint, hogy milyen szófajú névszót követhet. Ennek alapján lehetnek számnévi esetragok, főnévi ragok, melléknéviek. Ez a legtöbb esetben finomíthatja a beosztást.

Elvileg minden raghoz hozzárendelhetnénk egy-egy latinul megnevezett esetet, de ezt azért nem teszem, mert egy-egy raghoz több eset is tartozhat. Így pl. a *-nak* esetrag kifejezhet részes esetet, helyet (irányt) kifejező szót vagy birtokviszonyt. A *-t* rag általában a tárgyeset jele, de gyakran fejez ki mértéket, a *-val* pedig ugyanúgy lehet eszköz- és társhatározói toldalék. A morfológia jelentése eredendően alaktan, és most a jelentéssel csak másodlagosan foglalkozom.

Kihagytam azokat a ragokat, melyek már egyáltalán nincsenek használatban egy-két egyedi szót kivéve, mint pl.: *imént, rögvest*. Így a gyakorlatilag 32 (az alanyesettel együtt 33) esetet tartunk számon. Ezek közül 14 olyan van, melyet ritkán, vagy egyáltalán nem előzhet meg jel.

A személyes névmások ragozása is rendhagyó, de erről az 4.10.7 fejezetben foglalkozom.

4.10.3 Névszói jelek

A jelek kötött sorrendben előzik meg az esetragot. Szerepük alapján három csoportjuk van; egymáshoz képest előfordulási sorrendjük: birtokrag/többes szám, családi többes, birtokos jel,

Utolsó a **birtokos jel**, amely ha jelen van, jelölheti a birtok számát is: *Pistáéi*

nincs birtok	egyes szám	többes szám
	é	*éi*

Egy összetett toldalék, melyben az *i*-jel többes számot fejez ki.

Elöl áll a **birtok-** vagy **többesszám-jel**: *Pistája*/*Pisták* Összetett toldalék, mely (birtokos személyrag esetén) kifejezi a tulajdonos számát és személyét, de ezt kiegészítheti egy *-i* többes számjel, mely a tulajdon számára utal: *Kutyáink*, *fiai*.

Birtokos \ Birtok	egyes szám	többes szám
nincs birtokos		*k*
E.sz. 1. szem	*m*	*im*
E.sz. 2. szem	*d*	*id*
E.sz. 3. szem	*a*	*ai*
T.sz. 1. szem	*nk*	*ink*
T.sz. 2. szem	*tok*	*itok*
T.sz. 3. szem	*uk*	*ik*

A kettő között a **családi többes** (pluralis familiaris) állhat, ami nem minden névszóra alkalmazható: *bíróék*

Mivel az egyes toldalékok további jelekre való felbontása nehézkes (ezeknél a jeleknél az *i, k* többes számot jelöl, *m, n* első, *d, t* második személyt), ezért nem bontjuk tovább, hanem többfunkciós toldalékként tartjuk számon.

Megjegyzendő, hogy egyes nyelvészek szerint több birtokos jel is szerepelhet egy szóban (pl.: *fiáéé*), de a gyakorlatban ilyen formával még soha nem találkoztam. Szerintem egyfajta jel csak egyszer szerepelhet egy szóban.

További megszorítás lehet, hogy egy szóban nem szokott több többes számot kifejező jel lenni. Pontosabban, ha megvizsgáljuk az elmúlt évszázadok írásos műveit, nem találunk olyan szót, amelyben több többes számot kifejező jelet találunk, és ha mégis használnánk, kissé megüti a fülünket. Így pl. családi többes esetén mind az előtte, mind az utána álló jel csak egyes számú lehet: *fiaméké*
de kérdéses: ?*fiaimék* ?*Pistáékéi* **Pistáimékéi*

Ha ezt nem vesszük figyelembe, egy szóhoz 56 jelsorozat ragadhat, de az említett megszorítással is 28 különböző kombináció lehetséges.

Az, hogy a névszókat egy kalap alá veszem, az indokolja, hogy a toldalékolási korlátok nem olyan szófaji eredetűek, amely indokolná a szétválasztást. Például általános szabály, hogy számnévnek nincs többes száma. Ezzel szemben a következők természetes részei nyelvünknek: *sokak szerint, hármak bandája, az ötök…* Így általánosan állíthatjuk, hogy a névszói ragok és jelek 600 féle kombinációja használható minden névszónál,

4.10.4 Igeragozás

Az ige végén álló toldalék, amely meghatározza az ige módját, az ige alanyának számát és személyét, és utal az ige tárgyára is. A valóságban az igeragban egy – a finn nyelv morfológiájához hasonló – összetett toldalék nyomai fedezhetők fel:

igemód: kijelentő, felszólító, múlt, feltételes, infinitívuszi
tárgyasság: elkorcsosult utalás a tárgy kilétére
az alany személye: 1., 2., 3. személy
az alany száma: egyes és többes szám

Például: *látnálak*=lát+ n + ál + ε + ak
tő +feltételes mód+2.sz tárgy+1.sz+e.sz alany
láttunk=lát+ t + ε + un + k
tő +múlt mód+határozatlan tárgy+1.sz+t.sz alany

Míg a múlt idő jeléről gyakran beszélnek, az infinitívusz (főnévi igenévi) jelről ritkán, pedig formai szempontból nincs különbség.

Az igeragok ilyen jellegű szétszedése rendkívül bonyolult és a gyakorlatban felesleges, ezért azokat egy és oszthatatlan többfunkciós toldaléknak tekinthetjük. Persze nem minden esetben különböztethető meg a tárgyas és a tárgyatlan ragozás, és más esetekben is lehetséges, hogy különböző részjelek felszíni formája megegyezik: *nézzük* = felszólító/kijelentő mód
láttam = tárgyas/tárgyatlan ragozás

Az infinitívuszos ragok között szerepel olyan is, amikor nem fejezi ki sem az alany számát, sem a személyét: *tanulni*, de *tanulnom, tanulnod, tanulnia, tanulnunk, tanulnotok, tanulniuk*.

Igeragnak tekinthetjük a határozói igenév képzőjét is (*-va*/*-ve*, *-ván*/*-vén*), mivel ez is a szó végén szerepel, bár szerepe inkább az esteragokénak felel meg. Kérdés, hogy a *futtomban, futtodban, futtában* stb. alakok, valamint az *alhatnékomban, alhatnékodban, alhatnékjában* stb. alakok ebből a szempontból hasonló ragozott alakoknak tekinthetők, vagy egy képzett alak főnévi toldalékolt formáinak. Nem számolok a mai nyelvből kiveszett régies múlt idővel sem (*mondá az Úr*). Ezek nélkül is 60 körüli az igék ragozott alakjainak a száma.

A tárgyasság kifejezése a magyarban már elkorcsosult, szemben a finn és sok más finnugor nyelvvel, ahol az igerag a tárgy személyét, esetleg számát is kifejezi. Magyarban sem egyszerűen a tárgy határozottságát tükrözi a rag. A következő táblázat jól mutatja a cselekvő és a tárgy közti viszony kifejeződését.

Alany\Tárgy	indefinit	1. szem.	2. szem.	3. szem.
E.sz 1. sz	*nézek*	*nézem*	*nézlek*	*nézem*
E.sz 2. sz	*nézel*	*nézel*	*nézed*	*nézed*
E.sz 3. sz	*néz*	*néz*	*néz*	*nézi*
T.sz 1. sz	*nézünk*	*nézzük*	*nézünk*	*nézzük*
T.sz 2. sz	*néztek*	*néztek*	*nézitek*	*nézitek*
T.sz 3. sz	*néznek*	*néznek*	*néznek*	*nézik*

A táblázatból jól leolvasható, hogyha az alany személye magasabb, mint a tárgyé, akkor alanyi a ragozás, egyébként tárgyas. Ettől csak 1. személyű alany és 2. személyű tárgy esetén térünk el. Egyes számú alany esetén egy különleges ragot kap az ige, többes számú alany esetén alanyi a ragozás. Van olyan kivételes igénk, amelyben az egyes szám 3. személyű alany 1. és 2. személyű tárgy esetén nem a tárgyatlan igeragot kapja az ige, hanem ettől eltérőt. Ez az *eszik* ige. (*Engem esz a fene – kenyeret eszik – a kenyeret eszi*)

4.10.5 Képzők

A képzők azok a toldalékok, melyek a többi toldalékot megelőzik. Az a meghatározás, hogy a képzők azok a toldalékok, melyek megváltoztatják a szó jelentését (szemben a jelekkel, amelyek csak módosítják) elég homályos. Az igaz, hogy a képzők az esetek többségében megváltoztatják a szó szófaját. Főnévből melléknevet, igéből főnevet stb. képeznek. Azokat a toldalékokat is a képzők közé sorolom, melyek egy szófajból önmagára képeznek pl. *-hat* ige-ige-képző. Ellentétben a ragok és jelek használatával, a képzett szó teljes értékű marad, és újabb képzők csatolhatók hozzá.

A képzők lehetséges sorrendjét tulajdonképpen az határozza meg, hogy konzisztens-e az adott morféma-sorozat a szófaj-transzformációval. Persze ezen kívül sok más korlátozó tényező is szerepet játszik. Nem nagyon szerepelhet például egy toldalék kétszer egy szóban. Az egy szóban szereplő toldalékok száma is erősen korlátozott. Kérdés merül fel a magyar szavak szófaji felosztásánál is, hisz a szófaji határok átfedőek a névszókon belül. Gyakran használunk főnevet melléknévi szerepben, és viszont. Másrészt a hagyományos szófaji osztályokat érdemes tovább bontani, hogy pontosabb képzési szabályokat kapjunk. Ebből következően bizonyos toldalékok nem sorolhatóak egyértelműen az elvi kb. 100 csoportba. Némelyik több csoportba is beilleszthető, míg más csoportok üresek maradnak.

Bár a szófajokat még nem definiáltuk szabatoson, a képzők beosztása a következő lehet:

szófaj	aktív ige	passzív ige	főnév	fn/mn	tulajdonság	k.f. mn	tört-szn	sorszn	szelek-tor
aktív ige	*hat, gat, tat, *kál*	**dik, *kodik*	*ás, *alom, *mány, *at*	**ó*	*ó, tt, atlan, andó*				
passzív ige	*tat, *aszt*	*hat, gat*	**ás, *alom*		*ó, tt, atlan*				
főnév	*z, ít, *el, *kodik*	**ul, *ell, *dik, *zik, *dzik*	*ság, *ka, cska, *né*	**s, *ász, *ár*	*i, s, telen, ú, szerű, fajta, féle, kénti, *kori, *nkénti, *képpeni, *lagos, *beli*	*bb*			
tulajdonság	*z, ít, *kodik, *el*	*ul, dik, *zik, *dzik, *ell*	*ság, *ka, cska*		*s, talan, szerű, fajta, féle, kénti, kori, *nkénti, *lagos, *beli, *bani*	*bb*			
fn/mn	*z, *kodik*	**dik*	*ság, *ka, cska*		*s, talan, szerű, fajta, féle, nkénti, *nyi, *kori, kénti*	*bb*			
mérték-egység	*z, ít*		*ság, *ka, cska*		*s, talan, szerű, fajta, féle, nkénti, nyi, kori, *képpeni, *beli*	*bb*			
középf. mn	*ít*	**ul, *dik*	*ság*		*cska, szerű, fajta, féle, kénti, nkénti, *beli*				*ik*
tőszámnév	**el*	**ell*	*ság, cska*	**s*	*s, szerű, fajta, féle, nkénti, kori, szeres, szeri, rétű*	*bb*	*d*		
törtszámnév	*el*		*ság, cska*		*s, szeres, szeri, kori, nkénti, lagos, rétű*	**bb*		*ik*	*ik*
sorszámnév			*ság, cska*	*s*	*s*	*bb*			
szelektor	**z*		**ság, *cska*		*s, fajta, kénti, nkénti, *képpeni, féle*	*bb*			

A képzőkre sokkal nagyobb mértékben jellemző, mint a ragokra és jelekre, hogy a képző az egyik szóra alkalmazható a másik szóra nem. (Pl.: *jár-kál*, *buj-kál*, *mász-kál* stb. de nincs **híz-kál*, vagy **zár-kál*.) Az is fontos tény, hogy egyes képzők különböző szavak jelentését különbözőképpen módosítják. (Pl. *katonaság*, *asszonyság*, *hercegség*, stb.) Ha egy képzőről nem tudjuk megmondani valamilyen rövid és tömör formában, hogy melyik szóra alkalmazható, hogyan változtatja meg annak az értelmét, milyen lesz annak a szónak a ragozása, milyenek lesznek a keletkezett szó vonzatai, és nem csak az alaktannal foglalkozunk, akkor semmi okosabbat nem tehetünk, minthogy az összes ilyen képzett szót berakjuk szótárba, és ott adjuk meg a minden egyes képzett szó szükséges jellemzőit, vagy nem foglalkozunk a jelentésével, csak alakjukat vizsgáljuk. Van viszont egy sor olyan képző, aminél jól megadhatók a szón végbement változások; az ilyen képzőket hívjuk reguláris képzőknek. A csillaggal jelöltek nem regulárisak, de a csillaggal nem jelöltek sem minden esetben alkalmazhatóak. Általában egy végződést akkor tekintünk regulárisnak, ha meg tudjuk mondani milyen szavakhoz kapcsolható, és a keletkezett szó hogyan használható.

A ragokat és jeleket eleve reguláris végződéseknek tekinthetjük, noha mint láttuk, számos megszorítás lehet az alkalmazhatóságukra. Egy képzőt is akkor tekintünk regulárisnak, ha a megszorítások mértéke hasonlóan csekély, mint a ragoknál.

Mint jól látható, a középfok jelét én képzőnek tekintem, mivel akár más képző is követheti (pl.: *jobbító*). Azt, hogy milyen képzőt milyen másik követhet, alapvetően az határozza meg, hogy a képző milyen szófaj-transzformációt hajt végre. Egy biztos, az egy szóban található képzők számát a megérthetőség korlátozza. Így egy képző csak akkor ismétlődhet egy szóban, ha a korábbi képzett alak sajátos jelentéssel bír, így önálló képzetlen szónak tekinthető. Pl.: *biztonság-os-ság*. A képzősorozatok jelentése sem származtatható közvetlenül az egyes képzők jelentéséből. Pl. a *fut+hat+ó* képzősorozat nem lehet a *fut* ige alanyának jelzője, habár a *futó* igenév lehet.

4.10.6 Eltérések a sorrendi szabályoknál

Néhány esetben ragos alakok is kaphatnak képzőt. A *-ban, -ból, -kor, -nként, -szor, -képpen* rag után egy *-i* képző természetes, a *-szor*, *-lag* pedig *-s* képzővel folytatódhat. Ez felrúgja azt a szabályt, hogy a rag van a szó végén, de ha a *-bani, -beli, -kori, -nkénti, -szori, képpeni, -szoros*, *-lagos* képzőket felvesszük

toldaléktárunkba, akkor egy huszárvágással megoldottuk a gondot. Jobb, ha a szabály szabály marad, még akkor is, ha csak technikai megoldás egyszerűsíti a helyzetet.

A képzők sorrendjének szabályait lehet szigorítani, lazítani. Ha szigorítjuk, bonyolultabb nyelvtant kapunk, de kevesebb lesz a hibásan elfogadott (generált) alak, csökken a többértelműség esélye. Ha enyhítjük a szabályokat, egyszerűbb, kezelhetőbb modellt kapunk.

4.10.7 Egyéb toldalékok

A névszókon és az igéken kívül a névutók is kaphatnak toldalékot. ragozásuk hasonló, mint a névszók birtokos jellel való ellátása. Érdekes ezen kívül a személyes névmások többségének esetragos alakjainak előállítása, amely a ragok névutószerű ragozásával állítható elő. Egyes névutókból melléknévi névutó képezhető *-i* képzővel, mellyel hasonlóan viselkedik, mint a melléknév.

mellett, melletti, mellettem, melletted, mellette, mellettünk, mellettetek, mellettük
velem, veled, vele, velünk, veletek, velük

A személyragozott esetragokat és névutókat prefixként megelőzhetik a személyragnak megfelelő személyes névmás. A többes szám harmadik személy esetén viszont csak egyes szám harmadik személyű névmás szerepelhet: ***én****velem*, ***te****veled*, ***ő****vele*, ***mi****velünk*, ***ti****veletek*, ***ő****velük*.

Néhány idegen szavunk az eredeti nyelv alapján is toldalékolódhat. Pl. a latin *pejoratív, kumulatív*… szavainkból nem csak az *-an, -en*, hanem az *-e* rag is használható: *pejoratíve, kumulatíve*.
A legtöbb szó után lehet egy *-e* kérdőszócskát csatolni. Ezt nem toldaléknak, hanem **simulószó**nak, (**klitik**-nek) nevezik a nyelvészek. A magyar nyelvben csak a kérdőszócska az ilyen. Angolban az ige névmáshoz tapadó rövid alakját szokták klitikek közé sorolni. Oroszban a nyomatékosító *-ка* például ilyen. Tulajdonképpen csak definíció kérdése, hogy toldaléknak nevezzük-e vagy simulószónak. Formai, algoritmikus szempontból nem látok különbséget a kettő között.
Feladat 1: Írjon morfotaktikai CF szintaxist, amelyben a prefixek, a képzők és a simulószó kezelése megoldott. Mennyivel nő a grammatikai jelek száma, ha reguláris nyelvtannal kellene megoldani?

4.10.8 Előtagok és szóösszetételek

A magyar szavak elsősorban a végükön toldalékolódnak. Kétféle előtag szerepel a nyelvünkben, az igekötő, mely kizárólag igével járhat, és a felső, illetve a túlzófokot kifejező prefix, melyek középfokú melléknevekhez, illetve néhány úgynevezett pozicionáló melléknévhez kapcsolódhat. Ezen kívül szóalkotó eszközünk a szóösszetétel is.

Az igekötőt akkor kell egybeírni az igével, ha azt közvetlen megelőzi. Szótani szempontból csak ez az eset fontos. Az igekötők – hasonlóan a képzőkhöz – lehetnek regulárisak és esetlegesek. A *be, ki, le, fel, meg, el* igekötő szinte minden igéhez kapcsolható, míg a *helyre, rendbe, haza, agyon* stb. csak egyes igékhez. Azt, hogy mit tartsunk igekötőnek, ezek szerint magától az igétől is függ. Az irodalom szerint legalább 13 prefixet tartanak ebbe a csoportba, de ennél többet is sorolhatunk ide.

A felsőfok (*leg-*), és túlzófok (*legesleg-*) többnyire középfokú melléknevekhez kapcsolódhat, amitől az felsőfokú lesz. Egyes melléknevek, melyek általában szélső pozíciót jelentenek, alapfokú alakjaihoz is megkaphatják a fokozás eme morfémáját. (*legszélső, legelső, legutolsó, legfelső*…). Néhány szélső helyzetet jelentő határozószó, ragos főnév is kaphat melléknévfokozó prefixet, főként birtokraggal rendelkezők (*legalul, legtetején, legeslegszélére, legfenekén*…). Egy érdekes szabály: Ha egy középfokú melléknév megkapja a felsőfok prefixét, akkor abból már nem lehet további képzővel folytatni a szóalkotást: *szebb, szebbít, legszebb, *legszebbít.*

Az előtagok – szemben a toldalékok többségével – úgy kapcsolódnak a szóhoz, hogy sem az előtag, sem a szó alakja nem változik. Ebből a szempontból a szóösszetétel is ilyen kellemes szóalkotó eszköz. Annyi a különbség, hogy esetleg a szóalkotó elemeket kötőjellel választja el egymástól. Hogy mit lehet egybe, mit kell külön írni, az nyelvünk nem egyszerűen formalizálható része.

Ha a két számnév egyke sem összetett, egybeírjuk: *százhét*, *ezeregy*
Ha nem, akkor kétezerig egybe írjuk, ha ezer előtt nincs semmi: *száztizenkilenc*, *ezerkilencvennyolc*
Ha az ezer szó előtt szerepel valami, vagy kétezernél nagyobb, akkor kötőjellel választjuk el három számjegyenként a számot: *egyezer-egy*, *kétezer-húsz*, *háromszázegymillió-kétszázezer-százharminc*
Egyéb olyan összetétel, ahol engedélyezett a számmal való egybeírás, egyik tag sem lehet összetett: *sokemeletes*, *háromhavi*, de *tizenhárom havi*, *négy hónapos* (a *hónap* összetett szó)

Anyagneves összetételnél sem lehet egyik tag sem összetett: *rézdrót*, *vasakarat*, de *vörösréz színű*, *acél fogasléc.*

Általános szabály az úgynevezett hat-hármas szabály. Ha egy összetett szó három vagy több tagból áll nem számítva az egytagú igekötőket, és több mint hat szótagból áll a ragozatlan szó, akkor egy kötőjel használata kötelező: *tehergépkocsi-vezető*, *baleset-biztosítás, agy-érelmeszesedés*, *pirosfesték-keverő,* de: *balesetkárosult, baktériumkiiktatás, gyermekszótárírás.*

Még sok hasznos szabály van. Lehet a színekre, helységnevekre, intézményekre szabályokat találni, és más szemantikai megszorítást tartalmazó szabály létezik még. Ezeket cifrázzák még a mozgószabályok – mikor kell egybe írni egyébként külön írandó kapcsolatot, és fordítva – és persze vannak olyan egyedi esetek, amikor az általános szabálynak ellentmondó döntést hozott az akadémiai bizottság: *külügyminisztérium*, *belügyminisztérium*. Még egyszer hangsúlyozom, azokat a szabályokat, melyek algoritmizálhatók, használni kell, emiatt finomítani kell a szófaji osztályokat – anyagnév, mértékegység… – de sok esetben csak az egyedi minták döntenek. A *vajaskifli* egybe írandó (ha a vaj azt jelenti, hogy a tésztába keverik sütés előtt), a *mákos tészta* pedig külön. Miért? Bár néha létezik szemantikai magyarázat, de egyszerűbb azt mondani: csak!!!

Az összetételek megoldása nehéz feladatot ró az anyanyelvűnek is, de a számítógép számára sem lehet kielégítő algoritmust találni. Az csupán matematikai lehetőség, hogy soroljuk fel az összes lehetséges szóösszetételt. Mivel alapszavunk is számos van, ez a gyakorlatban megoldhatatlan feladat. Ezért a gyakorlatban a következő megoldásokat alkalmazzák:

1. A szokásos szótári tételek, még ha összetett szavak is, szerepelnek a szóellenőrző szótárban. Ezek feltehetően egyedi értelmezéssel bírnak. Ezek száma párszázezer, az adatbázisba vételük nem lehetetlen. Pl.: *őrvezető, kékfestő, vajaskifli, rendőr, őrnagy.*
2. Ha egyéb szabály nem tiltja – például a fenti 6-3-as – két szótári főnév összerakható: Pl.: *rendőrőrnagy, felügyelőbizottság, lábasztal, oroszláncsőr*. Emiatt bármely két nem összetett főnév egybeírható.
3. Szótári szavak képzett alakjára is megengedi az összetételt, ha azok a formák gyakoriak. Pl. tárgyas igéből képzett főnévi és folyamatos melléknévi igenév előtt bármilyen főnév állhat: *csonttöréskijelző, fásításszorgalmazás*.

Érdekes jelenség, hogy van pár olyan szavunk, mely önállóan nem szerepelhetnek, csak szóösszetétel első tagjaként. Ilyen sok rövid *o*-ra végződő szó, melyek hosszú *ó*-val szabályos szavak, de más a jelentésük: *fono, foto, sztereo, bio*… Egy másik csoportot alkotnak azok a szógyök, melyek képzett alakjait használjuk, de egyébként csak összetételbe, esetleg ragozott formában szabályosak: *gyógy, dörzs, ál, fek, össz, rak* (főnévként)…

Sok esetben nehéz dönteni: összetett, vagy egyszerű szóként kezeljünk valamit. *Visegrádot* kevesen érzik összetettnek, de *Leningrád*ot, *Volgográd*ot igen. *Szkóp*ból is sokféle van: *sztetoszkóp*, *oszcilloszkóp, kaleidoszkóp*, mégis nehéz a döntés. Az Akadémia álláspontja dodonai. Rábízza az író kompetenciájára. Hát igen. Én sem vagyok kompetens ilyen kérdésekben.

Feladat 1: Vegye számba, hány közös prefix módosíthatja a magyar névmást az angol *anybody*, *somebody*, *anyhow*, *someone*… mintára. Tehát melyek az alapnévmások, és melyek a módosító előtagok?

Feladat 2: Tesztelje: véletlenszerűen kiválasztott főneveket írjon egybe. Ezek hány százalékát tartja elfogadhatónak, illetve hány százaléka teljesen értelmetlen?

4.10.9 A toldalékok alakjainak számbavétele

Az előtagok és a szóösszetételek nem okoznak semmilyen változást a kapcsolódó morfémák alakjában. Ezzel szemben, még az angol nyelvben is láthattuk, hogy egy funkciót betöltő toldalék más-más alakban jelenhet meg, és eközben a szó, amelyhez kapcsolódik, maga is változhat. A magyar nyelvre ez még inkább jellemző. A toldalékok alakja (általában, de nem mindig) az őt megelőző morfémától függ. Eközben az a morféma is változtathatja az alakját, amelyhez kapcsolódik.

ε	*piros*
t, at, et, ot, öt,	*hajót, lovat, tehenet, tokot, tököt*
ban, ben	*pirosban, zöldben*
ba, be	*pirosba, zöldbe*
ból, ből	*pirosból, zödből*
n, an, en (on)	*forrón, pirosan, hidegen, gazadagon*
n, on, en, ön (an)	*hajón, lovon, tehenen, tökön, ? a pirosan*
ra, re	*pirosra, zöldre*
ról, ről	*pirosról, zöldről*

vá, vé *á, *é	*fává, kővé, sárrá, kenyérré*
nál, nél	*pirosnál, zöldnél*
hoz, hez, höz	*piroshoz, szekérhez, fűhöz*
tól, től	*pirostól, zöldtől*
ként:	*pirosként*
ért:	*pirosért*
nként, anként, enként, önként onként	*kettőnként, százanként, egyenként, ötönként, hatodonként*
kor	*háromkor*
ig	*háromig*
szor, szer, ször	*sokszor, tízszer, ötször*
szorta, szerte, szörte	*sokszorta, tízszerte, ötszörte*
szorra, szerre, szörre	*sokszorra, tízszerre, ötszörre*
val, vel *al, *el	*fával, kővel, sárral, kenyérrel*
nak, nek	*pirosnak, zöldnek*
ul, ül	*magyarul, törökül*
stul, stül, astul, estül, ostul, östül	*hajóstul, tetőstül, lovastul, tehenestül, botostul, tököstül*
képp(en)	*eredményképpen*
lag, leg	*alakilag, színleg*
onta, ente, anta, (önte, nta, nte,)	*naponta, hetente, nyaranta, ?őszönte, ?óránta, ?megyénte*
ott, ett, ött (t)	*Kaposvárott, Pécsett, Győrött, Vásárhelyt*
szerte	*országszerte*
szám(ra)	*zsákszám(ra)*
rét	*négyrét*
mód(on/ra)	*gombamód(ra)*

Jól látható, hogy a ragok alakjai attól függnek, hogy egyrészt kötőhanggal vagy a nélkül kapcsolódnak a szóhoz, másrészt a toldalék hangrendje lehet mély vagy magas, esetleg ezen belül lehet ajakkerekítéses és nem ajakkerekítéses. Elvileg ez nyolc lehetőség, de ezek között sok alakilag egybe esik, ezért hatnál több alakja egyiknek sincs. A kezelhetőségen belül marad. Kérdés lehet, hogy a kötőhang része-e a ragnak, vagy sem. A gyakorlatban célszerűebb annak tekinteni, mert az alakok száma ettől nem lényegesen nő, viszont a gyakorlati modellek jobban kezelnek olyan morfológiai leírást, ahol kevesebb összetevő van. Ez ellentétben lehet egyes nyelvészek modelljével, de mérnöki szempontból a modell lényege, hogy jól áttekinthető legyen az ember és a gép számára egyaránt.

Kiemelném a névszói teljes hasonulással járó *-vá/-vé* és *-val/-vel* ragokat, ahol a *v* betű helyett a ragot megelőző mássalhangzót kell duplázni. Ez a hangrendet is számításba véve elvileg 54 különböző alak, ami nem túl szerencsés. Itt, a korábbi indokok miatt el lehet térni a nyelvészetileg indokolt modelltől. Gyakori megoldás, hogy a szóvégi mássalhangzó-kettőzést a tő egy olyan alternánsának fogjuk fel, amely után e kérdéses toldalékok maradéka az alkalmazandó.

A névszói jelek, az igeragok és a képzők alakjai is ilyen értelemben felsorolhatók. A birtokjelnél az opcionális kötőhang a *j*, a magas-mély hangrendet figyelembe kell venni, de az ajakkerekítés nem játszik szerepet. Az igei toldalékoknak is több alakja is lehet, a szabályok összetettebbek. Pl. a kijelentő mód egyes szám második személyű rag lehet *-ol, -el, -öl* vagy *-asz, -esz, -sz*. Viszont az igékre is igaz, hogy egy (nyelvtani funkciót betöltő) toldaléknak soha sincs tíznél több megjelenési formája a magyar nyelvben.

Ha azt láttuk, hogy a toldaléknak nem egy alakja van, akkor azt is meg kell jegyezni, hogy attól függően, hogy milyen toldalékot használunk, az őt megelőző morféma is változtatja az alakját. Itt azért használom a morféma szót, mert a toldalék nem feltétlen közvetlen a szótőhöz kapcsolódik, hanem lehet, hogy egy képzőhöz, vagy jelhez. Ezt a változást úgy is fel lehet fogni, mint egy algoritmikusan jól megfogalmazott sztringmanipuláció, de azt is mondhatjuk, hogy egy morfémának nem csak attól függ az alakja, hogy azt milyen morféma előzi meg közvetlenül, hanem az is, hogy őt milyen morféma követ. Az első esetben a morféma ilyen jellegű változását akár futási (szóelemzési, szógenerálási) időben állíthatjuk elő. A más megoldás, amikor a lehetséges alternánsokat előre el kell készíteni, beosztályozni, így futásidőben kevesebb terhet ró az elemzőre. Megjegyzem, van olyan modell, amelyben a morfémák összes alakját algoritmikusan adják meg, vagyis, futásidőben generálódik a szükséges alak a jobb és a bal oldali környezet függvényében, így a toldalékoknak csak egy szótári alakja van (4.12.5).

4.10.10 Tőváltozások a magyar nyelvben

Ha a tő változásait formálisan akarjuk leírni, akkor a következő gyakori típusokat fedezhetjük fel:

Névszóhasonulás: *dob - dobbal*
Igei hasonulás: *húz - húzz*
Ige t-s átalakulás: *lát - lássuk*
Szóvégi magánhangzónyúlás: *kutya - kutyát*
Szóvégi v betű: *hó - havas, daru - darvak, nő - növő, fekszik - fekvő*
Szóvégi magánhangzó-rövidülés: *ajtó - ajtaja, tető - teteje*
Szóvégi magánhangzó-kiesés: *apa - apja*
Belső magánhangzó-kiesés: *pörög – pörgő, bokor - bokrot*
Belső magánhangzó-rövidülés: *szamár - szamarak*
Igei sz-kiesés: *eszik – ettek, fekszik - fekhet*
Harmadik azonos mássalhangzó-kiesés: *jobb - jobból*

Ezeknek a tőváltozásoknak számos alváltozata létezik, s egyesek egyszerre is előfordulhatnak: magánhangzó-kiesés+magánhangzó-rövidülés: *három – harmadik;* névszóhasonulás+harmadik mássalhangzó kiesése: *jobb – jobbal*; ige *sz*-kiesés+*t-s* átalakulás: *tetszik-tessék.* Az *sz* végű igéknél még gyakoriak az olyan tőváltozatok, melyeket kevésbé nevezhetünk tőtranszformációnak, inkább alternatív tőformának kell tekinteni: *alszik-aludni, törekszik-törekedni, mosakszik-mosakodni, alapszik-alapulni*, *haragszik-haragudni.*

Ha a szótári alakból akarjuk a tőalternánsokat előállítani, akkor néhány tőváltozás inverzét is fel kell venni. Ennek az az oka, hogy igéinknek nem feltétlen a szótő a szótári alakja, hanem az egyes szám harmadik személyű kijelentő mód, amelynél ikes igék esetén már lehet, hogy változást szenved a tő. Ebből a szempontból szerencsésebb lett volna az infinitívusz semleges alakja, mert az előbb említett *sz* végű igék és a hangzó-rövidülés kezelése egyszerűbb lenne. Más nyelveken az infinitívusz a szótári alak, de magyarban, mivel többnyire az igetövet adja az egyes szám 3. személy, eltérnek a nemzetközi gyakorlattól.

A tőváltozások között vannak olyanok, melyek hasonlóan értelmezhetőek. Például a belső magánhangzó-kiesés és a belső magánhangzó-rövidülés lényege, hogy a szóvégi mássalhangzó előtt, ha hosszú a magánhangzó, akkor vegyük a rövid párját, ha pedig rövid, akkor hagyjuk el. Ha írásunkban a finnek mintájára a megnyújtott magánhangzót a betű duplázásával jelölnénk, akkor az esetszétválasztás feleslegessé válna. Ebbe az osztályba vehetjük a csak néhány szóra alkalmazható betűátvetéses hangzókiesést is, mert minden szempontból azonosan viselkednek az ilyen szavak a hangzókiesésesekkel, ráadásul tisztán formailag a szótári alak alapján megállapítható, hogy betűcserés, vagy betűcsere nélküli a hangzókiesés: *teher–terhet.* Még egy általános jó tulajdonsága a magyar nyelvnek, hogy a tőváltozások mindig a szó végét érintik, emiatt nincs szükség az egész szó ismeretére, hanem csak szóvégi formális sztringátírási szabályokkal megadhatók. Ha pl. a németre gondolunk, akkor ott az umlaut alkalmazása nem a szó végén, hanem az első szótagon (pontosabban az összetett szó esetén az utolsó tag első magánhangzóján) alkalmazandó: *Großmutter – Großmüttern*

Más esetekben a tőtranszformáció nehezebben nevezhető meg, ezért inkább nevezik tőváltozatnak, mint tőváltozásnak. Algoritmikus szempontból nincs különbség.

A legtöbb változatú tövek az igéknél fordulnak elő. Régebbi igéink, mint az *eszik*, *alszik*, *lesz*, *jön*, *megy,* valamint a *tetszik, fekszik* típusú igéink változatosak ebből a szempontból

esz+ik	*tesz+em*	*feksz+ik*	*lesz+*	*jön+*
e+het	*te+het*	*fek+het*	*le+het*	*jö+het*
ev+ő	*tev+ő*	*fekv+ő*	*lev+ő*	*jöv+ő*
en+ni	*ten+ni*	*feküd+jön*	*len+ni*	*jő+ve*
é+ve	*té+ve*	*?feküv+ő*	*lé+vén*	*jer+*
egy+en	*tegy+en*		*legy+en*	*gyer+e*
	tév+ő		*lév+ő*	*jöj+jél*

A különböző modellek itt is eltérhetnek. Az igei teljes hasonulást (a felszólító mód toldaléka egyes igéknél nem *j*-vel, hanem betűkettőződéssel, esetleg *gy*-vel jelenik meg a felszínen) a toldalék különleges alakjaként, vagy a tő egy változataként lehet kezelni. Ha (gyakorlati okokból) a döntés ez utóbbi, akkor sem találunk olyan szót a magyarban, amelynek hétnél több alternánsa lenne.

Formálisan könnyű megadni a tőtranszformációkat. Az alábbi formalizmus a Helyeske szóellenőrző leírásában szerepel. Minden tőváltozástípust egy betűvel jelöltem esetleg egykarakteres prefixszel kiegészítve, melyet a lehetséges *minta-helyettesítés* pár listája követ.

A minta a szótári tő végét jelenti, melyet egy kötőjellel elválasztva a helyettesítendő karaktersorozat követ. Az ikes igéknél a szótári tő az *–ik* levágásával történik. Szerencsére nincs olyan iktelen igénk, ami *ik*-re végződne, így ez egyszerű karaktermanipulációval elvégezhető.

```
 0: -                                                                    # nincs változás
 k: --                                                                   # nincs változás, kötőjeles
 u: eksz-eküv, ugsz-ugov, sz-uv,                                         # igei v + beszúr
 v: ó-ov, ő-öv, új-úv, esz-év, isz-ív, sz-v, ud-v, üd-v, gy-v, -v        # igei v
 p: esz-é, isz-í, ön-ő, a-á, e-é, o-ó, ö-ő, i-í                          # tővégi magánhangzó nyúlás
 s: ió-iomo, árom-arma, ó-a, ö-e, a-, e-                                 # számnévi és főnévi magánhangzó rövidülés, kiesés
 r: ló-lov, u-v, ó-av, bő- bőv, ő-öv, ú-v, ü-v, nyű-üv, fű-füv, mű-műv, hű-hív, ű-v, hő-hev, é-ev, usz-v, üsz-v, esz-v, osz-v,
    asz-v, ehely-elyh, oholy-olyh, eher-erh, ohor-orh, ehem-emh, ohom-omh, lélek-lelk, a*-*, á*-a*, e*-*, é*-e*, i*-*, í*-i*,
    o*-*, ó*-o, ö*-*, ő*-ö, u*-*, ú*-u, ü*-*, ű*-ü                        # belső hangkiesés/rövidülés + v
 g: bb-bb, b-bb, ccs-ccs, cc-cc, cs-ccs, c-cc, ddzs-ddzs, ddz-ddz, dd-dd, dzs-ddzs, dz-ddz, d-dd, ff-ff, f-ff, ggy-ggy, gg-gg,
    gy-ggy, g-gg, th-tht, gh-ghg, ch-chh, hh-hh, h-hh, jj-jj, j-jj, kk-kk, k-kk, lly-lly, ll-ll, ly-lly, l-ll, mm-mm, m-mm, nny-nny,
    nn-nn, ny-nny, n-nn, pp-pp, p-pp, que-quek, qu-quk, q-qk, rr-rr, r-rr, zzs-zzs, ssz-ssz, zs-zzs, sz-ssz, ss-ss, s-ss,
    tty-tty, ty-tty, tt-tt, t-tt, vv-vv, v-vv, zz-zz, z-zz, x-xsz         # névszói teljes hasonulás
 h: b-b-b, ce-ce-sz, cs-cs-cs, c-c-c, dzs-dzs-dzs, dz-dz-dz, d-d-d, f-f-f, gy-gy-gy, g-g-g, th-th-t, gh-gh-g, h-h-h, j-j-j, k-k-k,
    ly-ly-ly, l-l-l, m-m-m, ny-ny-ny, n-n-n, p-p-p, que-que-k, qu-qu-k, q-q-k, r-r-r, zs-zs-zs, sz-sz-sz,  s-s-s, ty-ty-ty, t-t-t,
    v-v-v, z-z-z, x-x-sz, 1-1-gy, 2-2-v, 3-2-m, 4-4-gy, 5-5-t, 6-6-t, 7-7-t, 8-8-c, 9-9-c, 10-10-z, 20-20-sz, 30-30-c, 40-40-n,
    50-50-n, 60-60-n, 70-70-n, 80-80-n, 90-90-n, 000000-000000-v, 000-000-r, 00-00-z, 0-0-v, que-que-k,
                                                                         #névszói hasonulás, kötőjeles
 z: usz-, üsz-, sz-,                                                     # igei sz kiesés
 n: gy-n, sz-n                                                           # igei n betoldás
 o: *-o*                                                                 # hangzókiesés inverze igéknél
 ö: *-ö*                                                                 # hangzókiesés inverze igéknél
 d: eksz-eküd, aksz-akud, öksz-öküd, oksz-okud, elsz-elüd, egsz-egüd, agsz-agud, ögsz-ögüd, ogsz-ogud
                                                                         # ud, üd betoldás
-d: eksz-eked, aksz-akod, öksz-ököd, oksz-okod                            # od, öd betoldás
+d: szt-sz, sz-d, st-s, szt-s, t-s                                       # spec. felszólító mód hasonulása s-re, és iszik a d előtt
+s: t-ss, szt-ssz,t-ss,s-ss,tsz-ss
 l: van-val, -l                                                          # l vég
 j: jön-jöjj, megy-menj, úv-újj, hisz-higgy, esz-egy, isz-igy, eksz-eküdj, aksz-akudj, öksz-öküdj, oksz-okudj, elsz-elüdj,
    alsz-aludj, szt-ssz, sz-ssz, st-ss, dz-ddz, zz-zz, z-zz, s-ss, t-ts, -j                # a felszólító mód hasonulása + j
+j: jön-gyer, sz-gy                                                      # a felszólító mód hasonulása + j
 s: tsz-ss, t-ss, s-ss                                                   # a felszólító mód hasonulása s-re
```

A minták jobb oldalán (helyettesítési rész) lehet kötőjel. Ez csak névszói esetben fordul elő vagy változatlan tőnél, vagy teljes hasonulásnál. Pl. számjegyes számok toldalékolásánál. A mintákban a * egy speciális jel. Jelentése, hogy ott egy mássalhangzó áll. Ilyenkor a másik oldalon ugyanaz a betű helyettesítődik. Sajnos a típusok száma meghaladja a kognitív szép mennyiséget. Viszont jól olvasható, formális eszközökhöz nem szokott felhasználók is megértik, mégis precíz, jól algoritmizálható – véges fordítóval kivitelezhető. A minta-illesztésnél az első találatot fogadja el az algoritmus, de ha a leghosszabb egyezés az algoritmus alapelve, akkor annak is eleget tesz a leírás.

Ha az infinitívuszi forma lenne a kanonikus alak, akkor (az igei szótő a *-ni* levágása után értendő) egyszerűbb lenne a leírás, mivel a d, -d tőtranszformációra nincs szükség, az *sz* kiesés helyett pedig egy egyszerű *sz* betoldás kell:

```
 0: -                                                                    # nincs változás
 k: --                                                                   # nincs változás, kötőjeles
 v: ó-ov, ő-öv, új-úv, esz-év, isz-ív, sz-v, ud-v, üd-v, gy-v, -v        # igei v
 p: esz-é, isz-í, ön-ő, a-á, e-é, o-ó, ö-ő, i-í                          # tővégi magánhangzó nyúlás
+r: ió-iomo, árom-arma, ó-a, ö-e, a-, e-                                 # számnévi és főnévi magánhangzó rövidülés, kiesés
 r: ló-lov, u-v, ó-av, bő- bőv, ő-öv, ú-v, ü-v, nyű-üv, fű-füv, mű-műv, hű-hív, ű-v, hő-hev, é-ev, usz-v, üsz-v, esz-v, osz-v,
    asz-v, ehely-elyh, oholy-olyh, eher-erh, ohor-orh, ehem-emh, ohom-omh, lélek-lelk, a*-*, á*-a*, e*-*, é*-e*, i*-*, í*-i*,
    o*-*, ó*-o, ö*-*, ő*-ö, u*-*, ú*-u, ü*-*, ű*-ü                        # belső hangkiesés/rövidülés + v
 g: bb-bb, b-bb, ccs-ccs, cc-cc, cs-ccs, c-cc, ddzs-ddzs, ddz-ddz, dd-dd, dzs-ddzs, dz-ddz, d-dd, ff-ff, f-ff, ggy-ggy, gg-gg,
    gy-ggy, g-gg, th-tht, gh-ghg, ch-chh, hh-hh, h-hh, jj-jj, j-jj, kk-kk, k-kk, lly-lly, ll-ll, ly-lly, l-ll, mm-mm, m-mm, nny-nny,
    nn-nn, ny-nny, n-nn, pp-pp, p-pp, que-quek, qu-quk, q-qk, rr-rr, r-rr, zzs-zzs, ssz-ssz, zs-zzs, sz-ssz, ss-ss, s-ss,
    tty-tty, ty-tty, tt-tt, t-tt, vv-vv, v-vv, zz-zz, z-zz, x-xsz         # névszói teljes hasonulás
 h: b-b-b, ce-ce-sz, cs-cs-cs, c-c-c, dzs-dzs-dzs, dz-dz-dz, d-d-d, f-f-f, gy-gy-gy, g-g-g, th-th-t, gh-gh-g, h-h-h, j-j-j, k-k-k,
    ly-ly-ly, l-l-l, m-m-m, ny-ny-ny, n-n-n, p-p-p, que-que-k, qu-qu-k, q-q-k, r-r-r, zs-zs-zs, sz-sz-sz,  s-s-s, ty-ty-ty, t-t-t,
    v-v-v, z-z-z, x-x-sz, 1-1-gy, 2-2-v, 3-2-m, 4-4-gy, 5-5-t, 6-6-t, 7-7-t, 8-8-c, 9-9-c, 10-10-z, 20-20-sz, 30-30-c, 40-40-n,
    50-50-n, 60-60-n, 70-70-n, 80-80-n, 90-90-n, 000000-000000-v, 000-000-r, 00-00-z, 0-0-v, que-que-k,
                                                                         #névszói hasonulás, kötőjeles
```

```
 z: -sz                                                  # igei sz betoldás
+z: l-sz, d-sz, -sz                                      # igei sz betoldás 2
 n: gy-n, sz-n                                           # igei n betoldás
 o: *-o*                                                 # hangzókiesés inverze igéknél
 ö: *-ö*                                                 # hangzókiesés inverze igéknél
 d: eksz-eküd, aksz-akud, öksz-öküd, oksz-okud, elsz-elüd, egsz-egüd, agsz-agud, ögsz-ögüd, ogsz-ogud
                                                         # ud, üd betoldás
-d: eksz-eked, aksz-akod, öksz-ököd, oksz-okod            # od, öd betoldás
+d: szt-sz, sz-d, st-s, szt-s, t-s                       # spec. felszólító mód hasonulása s-re, és iszik a d előtt
+s: t-ss, szt-ssz,t-ss,s-ss,tsz-ss
 l: van-val, -l                                          # l vég
 j: jön-jöjj, megy-menj, úv-újj, hisz-higgy, esz-egy, isz-igy, eksz-eküdj, aksz-akudj, öksz-öküdj, oksz-okudj, elsz-elüdj,
    alsz-aludj, szt-ssz, sz-ssz, st-ss, dz-ddz, zz-zz, z-zz, s-ss, t-ts, -j      # a felszólító mód hasonulása + j
+j: jön-gyer, sz-gy                                      # a felszólító mód hasonulása + j
 s: tsz-ss, t-ss, s-ss                                   # a felszólító mód hasonulása s-re
```

Az Ispell típusú formalizmus (4.9 fejezet) ennél korszerűbb, viszont ragozási osztályonként külön le kell írni a mintákat és a helyettesítést, ráadásul a toldalékkal együtt. Itt viszont egy tőváltozási típust csak egyszer kell megadni, majd erre lehet utalni minden helyen, akkor is, ha egyébként a ragozásban különböznek a szavak. Ennél szebben, tömörebben Kimmo kétszintű morfológiája formalizál. (lásd: 4.12.5).

A tőváltozatok kezelése nem csak a szavakat, hanem a modellben minden olyan toldalékot érint, amelyet más toldalék is követhet. Ennek alapján a jeleknek és a képzőknek nem csak annak alapján lehetnek változatai, hogy mihez kapcsolódnak (tőlük balra milyen morféma áll), hanem attól is, hogy hozzájuk mi kapcsolódik az adott szóban (tőlük jobbra mi következik).

A szóösszetételeknél ilyen gond nem merül fel. Magyarban az összetett szavak csatlakozásánál semmilyen megkötés, változás nincs. A legtöbb, ami az úgynevezett harmadik azonos mássalhangzó tiltásának következménye, hogy esetleg a két összetevő kötőjellel kapcsolódik (*sakk-kör, balett-tanítás*). Az előtagok ebből a szempontból azonosan viselkednek a szóösszetétellel, tehát a csatlakozásnál nem változik sem az előtag (igekötő, felsőfok jele) sem a szó. Hogy ez is működhet másként, arra legyen példa a reform előtti német nyelv, melyben az összetett szavaknál működik a harmadik mássalhangzó elnyelődése:

Schiff+fahrt = Schiffahrt

4.10.11 Hangtani és egyéb illeszkedési szabályok (morfonológia)

Mint láttuk, a szavaknak is és az egyéb morfémáknak is több alakja van. Az, hogy melyiket lehet/kell az adott esetben használni, alapvetően hangtani szabályok határozzák meg. Ezek közül a legegyszerűbb, mely szerint magas hangrendű szavakhoz magas hangrendű toldalék, mélyhez mély dukál – már ha van választék *-ban, -ben*). Egy másik, hogy a toldalék ajakkerekítéses változatát kell alkalmazni, ha a szótő is az *(-hez, -höz).* Bonyolultabb szabályok határozzák meg az is, hogy pl. a birtokrag *j*-vel vagy a nélkül használandó (*tábornoka, boltja*). És még számos szabályszerűség befolyásolja, mely szavaknak milyen tőtranszformációja van, és melyik formájú toldalék illik hozzá. A *szamár* többes számban *szamarak*, a *tanár* pedig *tanárok*. Sajnos ezek a szabályok nem mindig következnek a szó(tő) alakjából. Még a legegyszerűbb kérdés eldöntése, hogy egy szó magas vagy mély hangrendű, sem minden esetben következik a szó alakjából. Az ökölszabály az, hogy ha összetett a szó, az utolsó tag számít. Ebben, ha minden magánhangzó magas, akkor a szó is magas hangrendű, ha pedig mindegyik mély, akkor a szó is az. Vegyes hangrendű szavaknál általában az utolsó magánhangzó dominál, de nem mindig. Nos, a gond az, hogy sem írásból sem beszédből nem derül ki, hogy az *i* magas, vagy mély. A *szív* szavunk magas hangrendű, ha főnév, és mély, ha ige (*szívunk, szívünk*). De ha a *kerék* szó magas hangrendű, a *derék* miért mély. Hát azért, mert származása a *dreko* szó, ami mély hangrendű. A *cél* szavunk a német *Ziel,* melyben mély hangrendű az *í*. Ezek miatt a szó eredetét is ismerni kéne. Időnként változik a szó besorolása. A Pisti – az István becézett alakja – valamikor mély hangrendű volt. Még Petőfi versében is így szerepel: „*Hát hogymint vagytok otthon, Pistikám*?"

A török és a finn nyelv hasonlóan bonyolult szószerkezettel rendelkezik, mint a magyar, de hangtani illeszkedési szabályai sokkal határozottabbak. A szótő alakja szinte mindig meghatározza, mely alakváltozat, mely toldalékkal kapcsolódik. Nem így a magyarban. A *ló*nak van *v*-s változata, a *vasaló*nak nincs. a *garázs* tárgyesete *garázst*, a *darázsé darazsat*. A *bokor*ból kieshet a második *o*, a *motor*ból nem. Habár erdélyi embereket hallottam már *motrot*, *garazst* mondani. Az *a*-ra, *e*-re végződő névszavaink teljesen egységesek, de egy angol nevet, franciás szót leírva lehet, hogy ki nem ejtett *e*-re végződnek, ezért kötőjelesen, és nem az *e* végűek

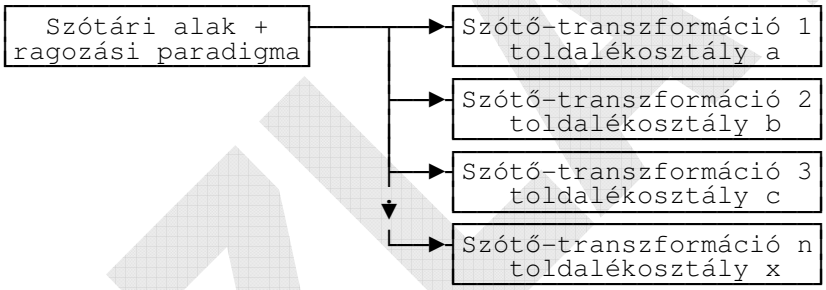
mintájára ragozódnak (*George-zsal, burlesque-kel*). Emiatt sok szónál, a szavak legalább negyedénél jelölni kell, milyen paradigmához tartozik, mert jelölés nélkül egyértelműen nem dönthető el a szótári alakból.

Lehetne a magyar nyelv leírásában is hagyatkozni a csupán szóalkotó karakterek alapján való besorolást, és ha egy szó ettől eltérően ragozódik, akkor azt kivételként kezelni. Az emberi feldolgozás szempontjából jó, ha a kivételek száma nem túl magas. Míg a török, a finn nyelvnél ez megoldható, a magyarnál a kivételek száma túl magas lenne. Emiatt a ragozást leíró formalizmusoknál célszerű, ha a szavak mellé nem csak azt kódoljuk, milyen szófajú, hanem azt is, milyen a viselkedése toldalékolásnál.

Jobb, ha a szavak (morfémák) szótári alakjához hozzá rendeljük az alakváltozatokat, és megmondjuk, azok milyen toldalékolási osztályba tartoznak. Ebben némiképp besegíthetünk számítógépes programmal, de a döntésnek emberinek kell lennie. Miután egy szótőhöz 10-nél kevesebb tőalak tartozik, ezt hasonlóan lehet készíteni, mint az angol teljes ragozási szótárát, vagyis összerendelni az alakokat az alapalakkal.

kutya → kutya(-, -ként ,-ság…,) kutyá(-t, -nak, -hoz, -val,-ja, -s, -é…)
ló → ló(-, -ként, -ság, -nak, -hoz, -val, -é…), lov(-at, -a, -as…)
madár → madár(-, -ként,-ság, -nak, -hoz, -é…), madar(-at, -a, -as…), madárr(al, -á…)
teher → teher(-, -ként, -ség, -nek, -hez, -é…), terh(-et, -e, -es…), teherr(-el. -é)
alapszik → alapsz(-om, -ol,-unk…), alapul(-, -hat,-ás, -ni…), alapv(-ó)
esz → esz(-ik, -em…), ev(-ett, -ő, -és…), e(-het,-ttem…), é(-ve, -én…), egy(-ek, -en…), en(-ni, -nem…)

A paradigmatikus osztályozásra épülő generatív modell ezek szerint a következő:



Itt a tőváltozások és a toldalékhalmazok függetlenül vannak definiálva, és bár nem függetlenek, tömörebb, tehát kezelhetőbb, jobban karban tartható leírást eredményeznek.

4.10.12 Az osztályozás módszerei

Mint korábban írtam, összetettebb szótannál sem a közvetlen alakfelsorolás (asszociációs módszer), sem az Ispellnél megadott módszer nem szerencsés. A nagymennyiségű adatok közt nem csak a felhasználók, de még a fejlesztők is eltévedhetnek. A hagyományos (leíró) nyelvészeknek erre az a bevált módszerük, hogy a szavakat osztályokba sorolják a ragozásuk szerint, amelyre egy mintaszóval megmutatják, hogyan kell toldalékolni, és a többi szónál erre, pontosabban az osztály elnevezésére utalnak. Ez két dologban tér el az Ispellnél használt leírástól.

1. Minden szó (kevés kivétellel) csak egy osztályba sorolódik
2. Az osztályok elnevezése utal egy fő- és azon belüli alosztály(ok)ra
3. Általában nincs formális leírás arra, hogyan kell ragozni – tedd a mintaszó alapján

Pl. egy ukrán szótár részlete:
бабах [isz]
бабахання {42}
бабахати {81} [foly]
бабахкання {42}
бабахкати {81} [foly]
бабахнути {85} [bef nép]
бабахнутися {85} [bef nép]
бабача {48-1}
бабаченя {48-1}
баба-яга {3A-1; 1Ґа}

egy magyaré:
buzi <1A1> *mn* ~{bb} és *fn*
buzog tn i <7a>
buzog|lány <4A>
buzogány||gyakorlat
bű[1] *se ~, se bá*
bű[2] **~***vöt v .~t; ~vök v.~ vek; ~ve*
bű||báj <2A6>
bűbájos. I.<15A>, II.<4A>
bűbájos|ság. <3A>
büdös I.<15C>, II.<2C4>

és egy spanyolé:
potencialidad *f*
potencializar [A4] *vt*
potencialmente *adv*
potencialomento *m*
potenciar [A1]
potenciómetro *m*
potentado -da *m,f*
potente *adj*
potestad *f*
potestativo -va *adj*

A spanyol esetén az igék okoznak gondot, ezért az igeragozásban adnak meg mintákat. A három nyelv közül a magyar a legnehezebb. A ragozási minta elnevezésében a kisbetű igét, a nagybetű névszót fed. Az első számjegy többnyire a kötőhangok létezésére, illetve tőváltozás módjára utal, és a betű a szó hangrendjét adja meg, a harmadik rész pedig az alosztályt jelenti. Néhány úgynevezett hiányosan vagy teljesen egyedien ragozható szónak a szótár megadja közvetlen az összes vagy néhány kritikus alakját.

Az ukrán példában az osztályok száma 100 körül van, ha nem számolom az alosztályokat. Az elnevezések vajmi keveset árulnak el arról, mit jelent az osztály. A spanyol igeosztályok sokkal kevesebben vannak, de itt sem árulkodik a név a ragozási típusról. Az Elekfi-féle osztályozásnál a kód függetlenül adja meg a hangrendi besorolást, az igei-névszói szétválasztást és a tőváltozás szerinti 20-nál kevesebb főosztályt. Ha beszámítjuk az alosztályokat, közel 600 típust lehet megkülönböztetni. Ha hasonló módszert szeretnénk finnre vagy törökre, ott is párszáz osztály lenne, de kevesebb, mint a magyarban.

Ilyen sok osztályt ennél praktikusabban kéne kezelni. Nem elég, ha a három ismérv szerint bontjuk szét a kódot, hanem a tőváltozást is nevesítenénk, esetleg osztályoznánk olyan résztulajdonságokra, melyek külön-külön kevesebben vannak, de egymástól logikailag független ismérvek.

Ennek tesz eleget a HUMOR leírás. Az eredeti (1993-as) formalizmus a morfémákat felszíni formáit sorolja fel szótáraiban. A felszíni formákhoz (formailag kigenerált morfémaalakokhoz) rendeli az elemi tulajdonságokat. A fő osztályozás mellett, mely alapvetően szófaji, illetve tőváltozat szerinti tulajdonságokat takarnak, illeszkedési tulajdonságokat kódoltak. Ezek elemi tulajdonságok, melyek többnyire binárisan eldönthetőek. Léteznek jobb és bal oldali elemi tulajdonságok. A jobb oldali tulajdonság a jobbról csatlakozó morfémák egyeztetésére szolgál, a baloldali pedig a baloldaliakéra. Az elemi tulajdonságokat 1-1 biten lehet reprezentálni. Ilyenek hogy magas vagy mély hangrendű a morféma, ajakkerekítéses vagy sem, a birtokrag *j*-vel vagy a nélkül csatlakozik, kell e kötőhang a toldaléknál vagy nem… A tulajdonságok persze nem ilyen egyszerűek. Például egy szóhoz egyes toldalékok kötőhangos, mások kötőhang nélkül, egyik toldalék ajakkerekítéses, másik nem (*füvön, füvet, tőrt, tőrön*). 16 elemi bináris tulajdonsággal írták le az illeszkedési osztályozást. A későbbiekben ugyan kiderült, hogy ez kevés, ezért a mai modell ennél többet használ, de használható formalizmus volt éveken keresztül. Két – egyéb szempontból összeillő – morféma akkor passzol, ha a bal oldalinak a jobb oldali, a jobb oldalinak pedig a bal oldali tulajdonsága megegyezik. Ez egy egyszerű összehasonlítással ellenőrizhető. Sok morfémánál mindegy, milyenek egyes tulajdonságok, ezért a tulajdonságok mellett egy maszkoló információ is van, ami arra jó, hogy pl. a bal oldal ajakkerekítésességét, a kötőhangi követelményt ne vegye figyelembe. Ez az algoritmust lényegesen nem bonyolítja. Az adatbázis egy részlete így néz ki:

ökörsütés	B 1101001011000011	~[FN]
ökörség	B 1101101011000010	~[FN]
öl	A 0110100000000001	~[IGE]
öl	B 1101101001010010	~[FN]
öl	B 1111101001010010	~[FN]
öldökl	E 011.111.1.10..0.	1öl[IGE]
öldöklés	B 1101001011000010	~[FN]
öldöklő	A 1110010010101010	~[MN]
öldököl	A 0110100000000000	~[IGE]
öldös	A 0110001010000100	~[IGE]
öleb	B 1101101011000011	~[FN]
ölel	A 0100100000000000	~[IGE]
ölelges	X 0100….1..0.10.	1t[IGE]

A példánkban szavak, nem toldalékok szerepelnek, ezért csak jobb oldali tulajdonságokat kell leírni. A tulajdonság első karaktere egy egyszerűsített folytatási osztályra utal, melyről későbbi fejezetben írok. Utána következnek a tulajdonsági bitek, melyek helyett időnként pont szerepel. A pont jelöli, hogy a szótő indifferens az adott tulajdonságra. Pl. az első bit választja szét az igéket a névszótól. A második mondja meg, hogy magas, vagy mély hangrendű a szó. A harmadik, hogy a toldaléknak, ami a szót követ, ajakkerekítéses, vagy a nem ajakkerekítéses változata használandó… A táblázat harmadik oszlopa a morféma – jelen esetben a szó – szótári alakjának helyreállítására szolgál.

Mostani szempontból az a lényeg, hogy a több száz osztályt tizenhét résztulajdonságra bontja, melyek különkülön ellenőrizhetőek, és melyek többsége bináris, de a betűvel jelölt sem vehet fel 28-nál több értéket. A legtöbb elkódolás persze itt fordult elő.

Ennek a formalizmusnak egyik nagy hibája, hogy nem a szótári alak az elsődleges, hanem csak származtatható. A mai HUMOR adatbázisnak nem is ez a kiinduló formája. Gondoljuk el, ha minden szótári tételnek átlagban két megjelenési formája van, akkor kétszer akkora adatbázist kell karbantartani, ráadásul a tételek nem függetlenek. Emiatt nő a hibázási lehetőség.

Kérdés, menyire lehet automatizálni a szavak besorolását. A gond ott van, hogy a tulajdonságok nem egyenes következményei a szó szótári alakjának. Mint korábban említettem, még a hangrendi besorolás sem következik a szó szótári alakjából. Ráadásul sok az ingadozó tulajdonság is. Például a tárgyeset *t*-je kötőhanggal és a nélkül is helyes a legtöbb *s* képzős melléknévnél. Anyanyelvén beszélő persze rávágja, hogy mely forma a helyes. Az igaz, hogy a hasonló végű szavak hasonlóan ragozódnak, de nem mindig. Az igéknél nincs gond az *-ít*, vagy az *-odik* végűekkel, de a *-z* végűek közt akad hangzókieséses is, a többség viszont nem az. Az osztályozásnál elengedhetetlen az emberi kontrol. Legfeljebb könnyíteni lehet a számítógép segítségével. A résztulajdonságok viszont nem teljesen függetlenek. Például, ha a tárgyeset kötőhanggal kapcsolódik a szóhoz, akkor a többi olyan toldalék is, melynek van kötőhangos változata. Több olyan „szakértői rendszer" készült, melyek lehetőleg kevés minta rákérdezésével segíti a szavak toldalékolási besorolását. Ezek között az a jó, mely összeveti a szó alakját a szótár eddig feltöltött elemeivel, és a valószínű tulajdonságokat ajánlja fel.

Ritka az a nyelv, melyben ekkora gondot jelent ez a feladat. Több nyelven próbálkoztunk automatikus besorolással. A francia például jó példa arra, hogy ne kelljen kézi beavatkozás a ragozási típus megállapítására. Talán a legkényesebb, melyhez mégis emberi döntés kell, hogy a szó eleji *h* hehezett, vagy sem. Ez a névelő egyeztetése szempontjából fontos. Németben sok szó neme következik a szó alakjából, míg másoké nem.

Folyamatos gond a típus kiválasztásánál, hogy a birtokragok *j*-s vagy *j*-tlen alakja fűződik a szóhoz. Ingadozó tulajdonság lehet a hangzókiesés is: *botolni-botlani*. szinte minden tulajdonságnál felmerülnek a döntés pontosságának a kérdése. Ezen a statisztika sem nagyon segít. És kivételek mindig vannak.

Pár érdekesség a magyar morfonológiához:

A *-ság*, *-ség* mindig a szó alapalakjához csatlakozik kötőhang nélkül. Kivéve a *gyorsaság*, illetve a *frisseség* szavaiknál. Bár ez utóbbinál, mivel sokan elírták, már megengedi a szabályzat a *frissesség* alakot, de ez alakilag nem a *friss* szóhoz, hanem a *frisses* szóhoz kapcsolja a főnévképzőt.

A középfok jele hitünk szerint a *-bb*, kivéve a *kevésbé* szavunkban.

A harmadik mássalhangzó elnyelődése jellegzetesen névszóknál fordul elő. Főként ha *-val*, *-vel* toldalék követi a duplázott mássalhangzóra végződő névszót. Nem is feltételeznénk, hogy létezik olyan igető, ami dupla mássalhangzóra végződik. De van, az *izzik*. Nos, itt például felszólító módban fellép a harmadik *z* elnyelődése.

Ha egy szónak van főnévi és melléknévi jelentése is, akkor a toldalékolásban különbözhet. Például a kötőhang ajakkerekítéses a főnévi jelentésnél, nem ajakkerekítéses a melléknévinél: *vörösök-vörösek*, *gyorsok-gyorsak*. Ennek a jelenségnek egy másik vetülete, hogy az általában melléknévhez kapcsolódó toldalékoknak akkor sincs ajakkerekítéses változata, ha egyébként a többi toldalék ajakkerekítésessel használandó. Ilyen az *-an*, *-en* módhatározó, és a középfok *-bb*-je. Ezek ugyan kapcsolódhatnak főnévhez is, de mássalhangzó után csak *a*, illetve *e* kötőhanggal. Azért itt is van kivétel: *gazdagon, szabadon*, *nagyon*, és ami a középfoknál teljesen egyedi: *nagyobb*. A *sok* középfoka pedig teljesen szabálytalan: *több*. Ez persze ne keserítsen el senkit. Az élő nyelvben mindig is vannak kivételek. De jó, ha a szabályok úgy fedik le a nyelvet, hogy a kivételek megszámlálhatók az ujjaimon. A kivételek kezelésére viszont minden formalizmusban legyen lehetőség.

Az *-i* képző *i*-re végződő főnevek után elnyelődik, mint a harmadik mássalhangzó. Emiatt a *férfi* szavunk melléknév is (lehet külön írni, hogy *férfi ruha*, akár a *női ruha* szavunkat), és nem kell a *hawaii*-ban megháromszorozni az utolsó magánhangzót, ha *Hawaii*ból származik valami. Ennek ellenére nem szerencsés *petőfi*-t írni, ha *Petőfitől valót* jelent

4.10.13 Morfoszintaxis – véges és végtelen modellek

A morfoszintaxis az egy szóban szereplő morfémák sorrendiségével foglakozik. Nem tér ki a morfémák alakjára. Tehát azzal foglalkozik, hogy igerag igetövet követhet, de névszóit nem, viszont hogy mi a konkrét alakja egy igeragnak, azzal nem. Ez az egyszerű toldalékoló nyelveknél nem gond. A szófaji beosztással

meghatározott az a nem túl nagyszámú toldalék, amit hozzátehetünk az alapszóhoz. A magyarban, törökben, finnben ez nem olyan triviális. Még az is kérdéses, hány lehetőség van.

Ha bárkit megkérdezünk, létezik-e 100 karakteres vagy annál hosszabb szó, mindenki azt mondaná, persze, hogy nincs. Ha esetleg valaki mégis kételkedik, akkor egy nagyobb számra már ő is az állítaná, hogy a szavak hossza korlátozható. Ha pedig így van, akkor matematikailag sem lehet végtelen sok szóalak egy nyelvben, hisz az abc rögzített. A legtöbb nyelv szótani modellje is ezt tükrözi. Az indoeurópai nyelvek leírásának szokásos módja, hogy megadják a szótárban szavak töveit, és azt, melyik ragozási osztályba tartoznak. Ezt vagy úgy, hogy egy-két toldalékolt alakját mutatják meg, melyekből a többi (emberi) logikával leszármaztatható, vagy úgy, hogy megadják a ragozási osztály kódját, vagyis utalnak egy leírás megfelelő elemére, amely megmondja, hogyan ragozandó a szó.

Ebből egy toldaléktár, egy szabályrendszer alapján, esetleg előtagtár segítségével felépíthetik az adott szó összes alakját, melyek mellé bekódolhatják a szóalakhoz rendelt összes nyelvi információt.

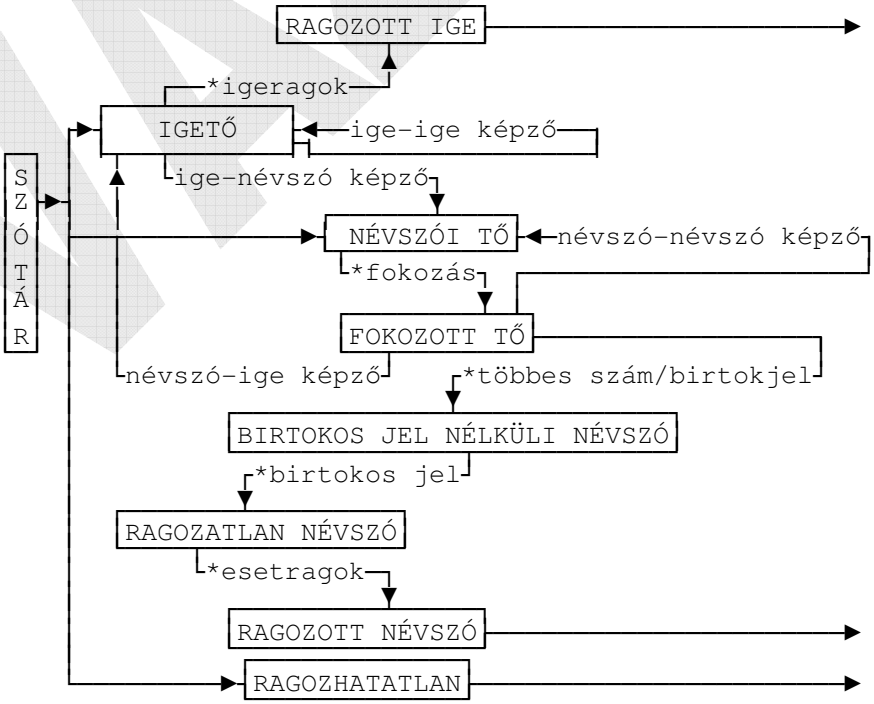
Ha ugyanezt akarjuk a korábban vázolt magyar morfológiai modellel végrehajtani, bajba kerülünk. A definíciók miatt a képzősorozatokon keresztül eljuthatunk egy szófaji kategóriából ugyanabba a kategóriába, pl. melléknévből melléknévbe juthatunk. Ha pedig így van, ki szabja meg, hogy ezt a képzősorozatot ne ismételjük meg.

káposztá+s+ít+hat+atlan+ít+hat+atlan+ít+hat+atlan+ít+hat+atlan+ul

Ha modellünkben le akarnánk tiltani az efféle esetet, akkor sokkal több osztályt kéne létrehozni. Bár a nyelvben valóban létezik mennyiségi korlát, de ez nehezen meghatározható. Az biztos, hogy a fenti példa esetén csak hosszas tűnődés után tudjuk megmondani, hogy a tányér tiszta, vagy káposztás. Hogy a megértés hányadik iterációnál válik nehézzé, az viszont az olvasó személyes tulajdonságától függ. Ha bevezetnénk mesterséges korlátot, mondjuk hármat, ami felé az ismétlések nem emelkedhetnek, akkor bár véges számú toldaléksorozat lesz csak elfogadható, de a nyelvi modell bonyolódik. Ha minden megengedett sorozatot külön tartunk számon, akkor több tízezer lehetőséget kell kezelni, ha pedig az ismétlések számát korlátozzuk, akkor pedig a képzők előfordulását kell elemzéskor számon tartani.

Ha nem tartunk a végtelen modelltől az iterációtól, egyszerűbb nem korlátozni a toldalékok számát. Lehet, hogy ezzel érthetetlen szavakat is beleveszünk a rendszerbe, de tudomásul kell venni. A szintaxis nem foglalkozik a jelentéssel, érthetőséggel. A szintaktikusan helyes mondatra nem szokás tiltásokat bevezetni akkor sem, ha értelmetlen. Miért kéne ez a szószinten megtenni?

Egy egyszerűsített példa egy 1990-ben készült szóelemző implementációjából:



Az átmenetek az állapotok között véges sztringhalmazok. Ha az átmenet *-gal jelölt, akkor akár üres (nulla hosszú) elem is lehet benne. A gráfban van ciklus is, de nincs ε ciklus. A prefixek és szóösszetételek kezelése ettől független séma alapján történik. Pl. a felsőfok prefixe akkor engedélyezett, ha szerepel a szóban *-bb* részsztring. Ez persze nagyon durva megközelítés, de mivel kevés olyan magyar szó van, amelyben szerepel *-bb*, de nem középfokú, nem okozott a gyakorlatban gondot. Mert ki ír le olyat, hogy *legrobban*. Vagy ha mégis, ez az ellenőrző szűrőjén átmegy.

A másik módszer, ha nem engedélyezzük a rekurziót. A magyarban a leggyakoribb megoldás, ha felsoroljuk az összes elfogadható képzősorozatot, és ezeket, mint összetett toldalékot használjuk. Ezzel a megoldással él a HUMOR és a HUNSPELL leírás is. A HUMOR-ban a képzői részre egy példa:

ésesebb	`]11011111.00.1010`	`m01.............`	és[IF]+es[SKEP]+ebb[FOK]
ésesed	`A01111010100000.0`	`m01.............`	és[IF]+es[SKEP]+edik[MI]

A HUNSPELL-nél a nyelvtani kategóriák adatai nem látszanak a toldalékolási szabályoknál:

```
SFX í 0 ésíthetett/400 [^aáeéiíoóöőuúüű]
SFX í 0 ésíthetetlenségileg [^aáeéiíoóöőuúüű]
```

Mivel az összes képzősorozat külön tételként fel van véve, kevésbé valószínű, hogy hibás sorozat elfogadtassék. Viszont nagy esély van arra, hogy egy-két ritkábban előforduló képzősorozat lemarad. Amikor azt tárgyaltuk, hogy hány alakja van egy szónak, akkor kiderült, több ezer használatos képzősorozat van a magyar nyelvben, és ha az alakváltozatokat is belevesszük, akkor 100 000 fölötti tételt kell kezelni ilyen esetben. Ez vetekszik egy normál szótár méretével. Mivel ekkora mennyiségnél a Zipf törvény következményeként előfordulhat, hogy ritkábban előforduló, de teljesen szabályos és érthető képzősorozat lemarad a listából, vagy hibásan kerül fel.

Ekkora szótárat nem szabad elemenként bővíteni. Mindkét implementációban generált állományról van szó. A szótárral szemben ez az információ nyelvre (pontosabban adott nyelvi modellre) jellemző, és rögzíteni lehet, mint a nyelvtan részét. A szavak tára viszont nyitott halmaz. Soha nem tudhatjuk, mi nincs benne, és nem lehet legenerálni tömörebb leírásból a toldaléktár mintájára.

4.10.14 A morfotaktika és morfonetika összekapcsolása

Míg a morfoszintaxis arról szól, hogy milyen sorrendben követhetik egymást a morfémák, a morfonetika – többnyire szomszédsági viszonyok alapján azt határozza meg, melyik felszíni formát öltik magukra az egyes morfémák. A morfoszintaxis blokkvázlatban megadott formája lényegében a nyelvtani szófajokkal operál. Emiatt annyi kategóriát (csomópontot) használ, ahány szófajt megkülönböztetünk. A magyar nyelvben a szófajok száma 10-20, a modell finomságától függően. A fonetikai tulajdonságok száma ennél nagyobb, de a legfinomabb beosztás szerint is csupán néhány száz. Elvileg, ha e kettő direktszorzatát nézzük, párezer kategóriával megússzuk. A kategóriák közötti átmenet a megfelelő morfémaalakok beolvasásával mehet végbe. Két kategória közötti közvetlen átmenetre csak véges sok morfémaalak segítségével lehetséges, tehát véges nyelvvel megadható. Egyes kategóriák elfogadható karakterláncot jelentenek, míg mások nem.

Tehát létezik véges sok kategória. A kategóriák egy része elfogadó, mások nem. A kategóriák közti átmenetet véges nyelvvel adjuk meg. Létezik egy vagy több kezdőkategória. Egy sztring akkor fogadtatik el, ha valamely kezdőkategóriából elindulva, a karakterlánc elejétől kezdve a következő részt úgy olvassuk be, hogy a beolvasott darab az a két kategóriát összekötő véges nyelv egy eleme, valamint így egy elfogadó kategóriába jutunk úgy, hogy a feldolgozandó karakterlánc végére jutunk, akkor a sémánk elfogadja a szót. Az ilyen definíció reguláris nyelvet definiál.

Ha tehát csupán az a célunk, hogy megállapítsuk, egy sztring a nyelv szavaihoz tartozik-e, akkor ezt véges automatával ellenőrizni lehet. Ezt az automatát a paradigmatikus leírás alapján közvetlenül elő lehet állítani. Az Elekfi-féle leírásban mintegy 600 paradigma szerepel. Minden paradigma – a tőalternánsoknak megfelelően – 1-7 folytatási osztályt határoznak meg. Szerencsére sok egybeesés van. Különböző paradigma egyes folytatási osztályai azonosak. Emiatt a folytatási osztályok száma is 400 alatt marad. Ha pontosabb modellt akarunk, akkor párezer kategória biztosan elegendő.

A HUNSPELL leírásából nehezebb, de azért elő lehet állítani hasonló sémát. A HUMOR formalizmus is alkalmas arra, hogy véges automatával reprezentáljuk. Itt a fonetikai tulajdonságokat kell összeolvasztani. A becslések szerint a ténylegesen előforduló osztályok száma itt is párezerre tehető.

4.11 Morfológiai eszközök

4.11.1 Elemzés, ellenőrzés, generálás

A szótan formális leírásának célja, hogy belőle elemzőt, generátort, ellenőrzőt, szövegkorrektort készítsünk. A nyelvi modellek többnyire generatív módon vannak megadva. Tehát a leírás elsődlegesen arra szolgál, hogyha megadnak egy szót és hozzáteszik a bővítés nyelvi információit, akkor ebből megkaphassuk a szó aktuális formáját.

Ha úgy tekintjük, hogy az elemzés egy karaktersorozathoz (a felszíni alakhoz) egy olyan karaktersorozatot rendel, amely az elemzést reprezentálja (a morféma szótári formája + nyelvi információ és ezek sorozata) adja, akkor matematikai értelemben fordítóról van szó.

Az asszociációs módszernél akár elemző, akár generatív leírás a forrás, a másikat a reláció megfordításával könnyen megkaphatjuk. Miután minden egyes szóhoz (vagy szóalakhoz) egyedi információt kell rendelni, a legtakarékosabb tárolás a betűfának egy olyan változata, amely a kulcshoz hozzárendel egy egyedi számot (pl. a sorszámát). Így az összerendelés, az asszociáció természetes módon, a sorszámmal kódolható. Fák módosítása egyszerű, emiatt dinamikusan változó szótárak használatánál is bevált. Ha nincs szükség elemzésre, csupán szóellenőrzésre, akkor egy véges nyelv tárolására az őt reprezentáló minimálautomatára van szükség, ami véges nyelv esetén (rekurzió-) ciklusmentes gráffal adható meg. Az automata determinisztikus, emiatt az ellenőrzés sebessége ténylegesen az input hosszával lineáris időben elvégezhető. Ennek módosítása, akár egy új szóalak hozzáadása viszont nem egyszerű feladat, gyakorlatilag elölről kell generálni az automatát. Egy minimálautomata generálásának időigénye elvileg az állapotok számának exponenciális függvénye (a determinisztikusság követel sok időt), ami nem szerencsés egy dinamikusan változó szótár kezelésére. Gyors algoritmus létezik előállítására. Ez a Frey algoritmus. (4.12.2)

A konkatenációs módszer is, ahogy az előző fejezetben láthattuk, véges automatával reprezentálható közvetlenül. A gond ott van, hogy egyes osztályok közötti átmenet determinisztikussága könnyen biztosítható, hisz véges nyelvvel van definiálva, de az már nem, hogy kiválasszuk, mely osztályok közötti átmenetet óhajtjuk meglépni. Elvileg persze minden véges állapotú automatának egyszerű algoritmussal megszerkeszthető a vele ekvivalens determinisztikus, de magyar nyelv esetén az állapotok száma is milliárdokat tesz ki, tehát technikailag lehetetlen. Emiatt az implementációkban ezt elkerülik. Ehelyett több automata paralel futtatását választják (reguláris nyelvek zártak az unióra és a metszetre), vagy többszintű automatát alkalmaznak, vagyis az inputot véges fordítóval transzformálják, és az eredményt véges állapotú automatával ellenőrzik. A reguláris nyelvek osztálya ezekre a műveletekre zárt.

Ha nem egyszerű ellenőrzésről van szó, hanem a szó elemzése is fontos, akkor véges fordítóval ez is megtehető. A fenti modell mindegyike alkalmas arra, hogy véges fordító kiadja a morfémák szótári alakját az esetleges nyelvtani információval együtt. Az alaplogika az, hogy olvasás közben tárolva a kiinduló osztályt, a beolvasott morfémaalakot, konstatálva a morfémaalak végén elért osztályt – ezek véges lehetőséget adnak – tehát hozzárendelhetünk bármilyen sztringet outputként. Persze az implementációk nem követelnek ennyi megjegyeznivalót, nem is lenne érdemes, mert akkor az állapotok száma nagyon nagy lenne. Itt viszont már reménytelen determinisztikusságot várni. Már elméletileg sem lehetséges, hisz egy szó(alak)nak több elemzése lehetséges. Márpedig ha egy függvény nem egyértelmű, akkor azt nem lehet determinisztikus eszközzel megvalósítani.

A generálás elvileg szintén véges fordítóval megvalósítható. Mivel a nyelvi modellünkben az ε morfémák (azok, melyek hossza 0) nem léphetnek fel ciklusban, a véges fordító funkciójában megfordítható. Persze ebben az irányban is létezik többértékűség, ezért itt sem reménykedhetünk determinisztikusságban.

A különböző céloknak lehet közös az adatbázisa, de el is térhet. Az ellenőrző dolga, hogy megállapítsa, vajon van-e helyes elemzése a karakterfüzérnek. Ebből ha egyet talál, már el is végezte a feladatát. Az elemző ráadásul megállapítja a lehetséges elemzéseket. Ha lehet, mindegyiket. A generátornak nem szükséges az összes felszíni megvalósítást kiadnia, elég egyet létrehoznia. E három funkció miatt az adatbázisuk is eltérhet. Ha generálunk, nem kell minden alternatívát megadni. Elég a dominánst, ha van. Azt is mondhatnánk, hogy a gépnek is lehet a morfológiában egy aktív és egy passzív nyelve, Passzív, amit elfogad, elemez, aktív, amit generálni képes. Minek a ritkán hallott *aluszom*, mikor a gyakori *alszom* is jó, ha generálunk. Elemzésnél persze mindkettőre fel kell készülni. A passzív nyelv a gép esetében is bővebb, mint az aktív. Az szótani leírások nem kis része **szimmetrikus** – a két nyelv megegyezik, tehát nincs különbség a generált és az elemzett nyelv között.

4.11.2 Többértelműség a szavak szintjén

A természetes nyelveknél – szemben a mesterséges programnyelvekkel – elkerülhetetlen a többértelműség. Az eszperantó nyelv tervezése folyamán ugyan törekedtek az egyértelműségre, de ez nem sikerülhetett. Ha egy nyelv élni kezdi a saját életét, természetes módon jelentkezik az értelmezés gondja.

Ha futó szövegekben megnézzük, a szavak hanyadrésze elemezhető többféleképpen, akkor különböző nyelvek más-más arányt mutatnak. Érdekes, azt mutatják a felmérések, hogy az angol és a magyar hasonlóan többértelmű a szavak szintjén. Körülbelül minden harmadik szóalak elemzése nem egyértelmű. Persze kérdés, mit jelent a többértelműség. Angolban a többértelműség leggyakoribb oka, hogy egy szóról önmagában nem derül ki, névszóról, vagy igéről van szó. A *black feketét* jelent, de jelenthet *feketít*-et is. A *file* egyik jelentése *ráspoly* – mint ige *reszel*, a másik pedig *állomány*-t jelent vagy igeként *fájlba vételt*. Persze előfordul más többértelműség is.

A ragozó, de nem sok szóalakot használó nyelvekben kevesebb a többértelműség. A franciák mindig az *avions* (igeként a birtoklás t. sz. 1. szem. alakja, főnévként repülőgépeket jelent) szóval mutatják, hogy rendszerük képes a probléma kezelésére. Nincs sok más példájuk.

A magyarban is léteznek olyan szótári szavak, melyek alakjukban megegyeznek, de semmi közük egymáshoz: *lép,* amit a méhkaptárban találunk, *lép*, ami a máj melletti szervünk, és *lép* igénk is van. Nem ebből ered a nagyszámú többértelműség. A gond ott van, hogy a toldalékok önmagukban is lehetnek többértelműek. Ha azt mondom, hogy „*Felkerestem a falu bölcsét, hogy kikérjem a tanácsát – a bölcsét”*, akkor a két azonos alakban az *é* jel teljesen más funkciót tölt be. *Én fagylaltot szeretnék, de szüleim azt szeretnék, ha nem ennék*. Itt is mindenki tudja, hogy a *szeretnék* szónak mikor mi az alanya, de az önálló szóból ez nem derül ki. Aztán vannak olyan esetek, amikor egy szóalaknak meglepően sok elemzése van. Lehetséges, hogy a szótő is, a rag is többértelmű, bár azt lehet tudni, melyikhez melyik tartozik: *A várvédők segítséget nem várnak, a várnak el kell esnie*. Aztán vannak olyan esetek, amikor az ember nem is gondol a sokféle elemzésre, mert az egyik – bár a nyelvi modell szerint szabályos – az életben senki sem használja: *malachit*ot mondunk a kőzetről beszélve, de *malachit*et kell használnunk, ha a disznók hitéről van szó.

Hogy mitől vannak az egyes többértelműségek, azon kár elmélkedni. Hogy mitől ilyen nagyszámú a magyarban a többértelműség, annak az az oka, hogy toldalékolás és szóösszetételek miatt a magyar nyelv szavai nagyon **sűrű**n vannak. Ezen azt értem, hogyha véletlen generálunk mondjuk hatkarakteres betűsorozatokat, akkor magyarban nagyobb az esélye, hogy értelmes szót kapunk, mint például csehben vagy franciában. A szavak formailag többen vannak „ugyanabban a térben”, így időnként két különböző szó egybe is eshet.

Ha a szavak egyharmada legalább kétértelmű, akkor körülbelül kilencede három elemzéssel bír, huszonheted része néggyel… Persze a statisztika nem ilyen pontos. Ritka a tíz különböző elemzéssel bíró szó.

Íme, egy sokjelentésű szó: *értem* a HUMOR elemzése szerint.

én[NM]=+ért[CAU]+=em
ért[IGE]+em[Te1]
érik[IGE]=ér+tem[Me1]
érik[IGE]=ér+tem[TMe1]
érik[IGE]=ér+t[MIB]+em[PSe1]
ér[IGE]+tem[Me1]
ér[IGE]+tem[TMe1]
ér[IGE]+t[MIB]+em[PSe1]

Ebből a nyolc elemzésből kettő természetes következménye a magyar ragok többértelműségének: Múlt idő egyes szám első személyben nem lehet megkülönböztetni a tárgyas és tárgyatlan ragozást. E nélkül is marad hat lényegesen eltérő elemzés.

A többértelműség száma függ a nyelvi modelltől és az aktuális adatbázistól. Ha finomabb, részletesebb a szó nyelvtana, akkor sok eset eshet egybe. Például, ha megkülönböztetjük a főnévi *-s* képzőt az melléknévitől, a főnévi *-ó*, *-ő* képzőt a melléknévitől, a legtöbb esetben duplázódik az elemzés. Ha viszont az előbb említett múlt idejű tárgyas és tárgyatlan ragozást összevonjuk, akkor csökken ezekben az esetekben a kötelező kettős elemzés. A szótár milyensége is befolyásoló tényező. Ha szerepeltetnünk kell összetett szavakat – mert például szótár jellegű alkalmazásról van szó – akkor a szó egyben, és alkotó elemeire szétszedve is megjelenhet elemzésként.

A többértelműség sok esetben zavaró lehet. Emiatt azokban a nyelvekben, ahol ez nagyszámú, különböző módszerekkel próbálják feloldani, csökkenteni számukat. Ennek fontos jelentősége van az elválasztó programoknál is, de a mondatelemzők, fordítók esetén is hasznosak. A generátorok esetén általában szűkítik a nyelvi szabadságot, hisz majdnem lényegtelen, melyik helyes felszíni formával reprezentáljuk a szót. Emiatt a generátorok akár egyértelművé, ezen keresztül determinisztikussá is tehetők. Más esetekben való-

színűségi módszereket használnak. Ha egy elemzés valószínűsége lényegesen kisebb, mint a többi, akkor ezt elvethetjük.

Az angol nyelvben, bár a többértelműség hasonló relatív gyakorisággal fordul elő, mint a magyarban, a vizsgálat akár a teljes szóállományra elvégezhető. A szóalakok száma felsorolható, és a mai számítástechnikai eszközökkel ezekből a többértelmű szavak lekérdezhetőek. A munka nem egyszerű, mert többnyire emberi döntéssel kell megállapítani, hogy a szónak az adott szövegkörnyezetben melyik értelmezése érvényesül.

A magyar nyelvben a többértelműség ilyen szűrése nem kivitelezhető. Még a többértelmű szavak felsorolása is lehetetlen. Ennek ellenére lényegesen csökkenthető a többértelműség. A módszer lényege, hogy a gyakori többértelműség okaira kell fényt deríteni. Hogy mi a gyakori? Erre ad választ a statisztika.

1. Keressük ki a leggyakrabban használt 10 000 szóalakot, és nézzük meg, melyeknek van különböző elemzése. (kb. 3000 szó lehet). Ezekről döntsünk, és ha érdemes, tiltsuk le azokat az elemzéseket, melyek egyáltalán nem fordulnak elő az életben. Például:

   egyelőre[HA] kosár[FN]
   egyel[IGE]+*ő*[MIF]+*re*[SUB] *kos* [FN]+*ár*[FN]

2. Nézzük meg, a gyakori nem várt elemzések között van-e közös ok, mely általános módszerrel küszöbölhető ki. például sok *-ok* többes számú névszó összetett szóként is értelmezhető:

   bolond[FN]+*ok*[PL]
   bolond [FN]+*ok*[FN]

Ilyen eseteket például úgy lehet megszüntetni, hogy az *ok* főnevet nem engedjük meg algoritmikus szóösszetétel második tagjaként (ha pontosabb eszközünk van, a mélyhangrendű *o* kötőhangos főnevek után), illetve azokat az összetett szavakat, melyekben mégis előfordul, külön szótári tételként felvesszük.

Egyértelműsítésnek még számos eszköze van. Lehet szövegkörnyezet figyelésével nyelvtan, illetve valószínűség alapján dönteni, de ezek a módszerek túllépnek a szóhatáron.

Nyelvtani szövegkörnyezet-vizsgálattal gyakran megállapítható, hogy az *az* nem lehet névelő. Például ha egy ige, kötőszó vagy névutó követi, vagy egy olyan névszó, mely (kiejtésben) mássalhangzóval kezdődik, akkor eleve csak vonatkozó névmásról lehet szó.

A statisztikai módszereknél többnyire szófaji n-grammok módszere válik be. Ha szóalakok hármasainak eloszlását néznénk csak a sokszor előforduló szavakra, akkor is kis biztonsággal tudunk döntést hozni.

4.12 Implementációs módszerek

4.12.1 Heurisztikus módszer alkalmazása magyar szóelemzésre

Korábban bemutattam egy heurisztikus elemzőt, mely angol nyelvre működik. Az ilyen egyszerűsített módszereknek inkább akkor van jelentősége, ha

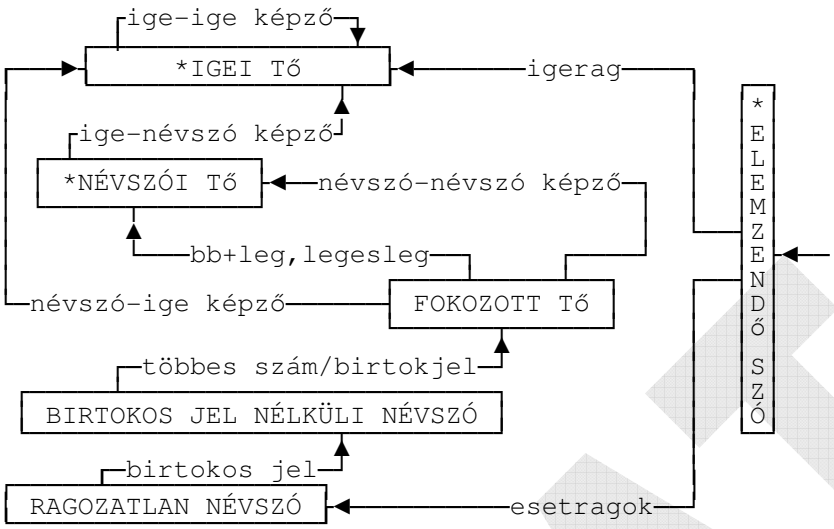
1. Nem kellően feltérképezett nyelvről van szó
2. Ismeretlen szavak elemzése szükséges

Az angol nyelv toldalékolása elég egyszerű ahhoz, hogy a heurisztikus módszerrel 99 százaléknál jobb lefedettséget biztosítson, és a hamis elemzések száma sem számottevő. A következő példa abból az időből származik, amikor még nem létezett használható digitalizált szótani leírása a magyarnak. A módszer lényege, hogy a szavak szótári alakját és szófaját tartalmazó szótárból és egy aránylag pontos toldaléktárból állt. A szótár ilyen módon laikus segítségével is építhető. A toldalékalakok tára kis munkával előállítható. Tehát két szótár létezik, a tőtár és a toldaléktár. Az elemzés lényege hasonló az angol példabelihez:

1. Nézzük meg, szerepel-e szó a szótárban. Ha igen, megtaláltunk egy elemzést.
2. A toldaléktár alapján próbáljunk levágni egy toldalékot a szóról. A vágás helyén ellenőrizzünk minimális fonetikai illeszkedést, és ha kell, próbáljuk rekonstruálni a toldalékra jellemző tőváltozást. A maradékra kezdjük újra az algoritmust.

Tehát alapvetően a szó végéről igyekszünk levágni toldalékokat, és a maradékot a tőtárban keressük. A szótárban a szavak szófaja is szerepel, míg a toldaléktár a morféma funkcióján kívül némi információt tartalmaz a fonetikai illeszkedésről. Ez a leírás, szemben a részletes Elekfi szótárral, mindössze 32 osztályt használ, amiben a toldalékot megelőző lehetséges karakterre szorítkozik, és a lehetséges tőtranszformációt adja meg.

A morfotaktikus gráf a 4.10.13 fejezetben említetthez hasonló Ennek is természetesen a megfordítottját, az elemzési gráfot kell alkalmazni. A csillaggal jelölt állapotokban lehet szótári keresést végrehajtani:



A toldaléktárból egy részlet – a névszói esetragok kódolása. A ragformákat megelőző szám a fonetikai osztályt jelenti. A ragot követő szám a toldalék kódolása:

```
1:-           : 0          2:ig                                :61
1:ként        :50          2:szor,ször,szer                    :64
1:kor         :60          2:nak, nek                          :71
2:ban, ben    :20          2:ul, ül                            :72
2:ba, be      :21         13:t, 25:at,et, 5:ot,öt, 9:vat, vet  :11
2:ból, ből    :22         10:n, 25:en                :38=30 vagy 34
2:ra, re      :31               25:an                         :34
2:ról, ről    :32                        5:on,ön, 9:von, vön  :30
2:nál, nél    :40         19:á,é, 10:vá,vé                    :35
2:hoz, hez,höz:41         10:nként, 5:anként,enként,onként,önként :54
2:tól, től    :42         19:al,el,7:val,vel                  :70
2:ért         :51
```

Például az `1`-es fonetikai kód tőváltozás nélküli korlátozásmentes illeszkedést jelent. A `2`-esnél, ha hosszú magánhangzóra végződik a maradék, lehetséges annak rövidítése. A `9`-es előtt csak magánhangzó szerepelhet, és megnyújtható. A többértelmű toldalékalak speciális kódot kap: `n`, `en :38`, ami vagy `30` (helyrag), vagy `34` (módhatározó).

Természetes ez a kódolás nem minden lehetőséget ír le. A ritka, speciális tőváltozatokat, a különleges szóformákat közvetlen bevehetjük a szótárba. Egy részlet a tőtárból:

```
terh, 21006:teher|21004
terhes, 22014
tevőd, 10005
téesz, 21004
téged, 06206:te|06204:0et011
tégedet, 06206:te|06204:0t011
```

A szó utáni kód szófaji besorolást és esetleges elemzési segédinformációt jelent. Ha: követi a szót, akkor annak valamely kifejtése szerepel utána. Ezt az algoritmikusan nem kezelt kivételek kezelésére használtuk. Például a *terh* tőforma után a szótári alak szerepel. A *téged* szónak nyelvtani felbontását adtuk meg.

A leírást szóelemzés céljából készült. Feltételezve, hogy helyes szöveget kap az elemző, minden elemzését előállítja. Mivel a fonológiai illeszkedés elég durva – a szavak alakjából következtet csak a lehetséges osztályra – többször fordul elő túlelemzés. Például a *Béla* mondat elején akár a bél szó birtokragos alakja is lehet, mert az *é* hang nem feltétlen jelent magas hangrendű szót. A tapasztalatok szerint az elemző 20 százalékkal több elemzést ad, mint egy precízebb nyelvtani rendszer. Abban az időben viszont nem volt más. Ma is ajánlom ezt a robusztus módszert fel nem térképezett nyelvek feldolgozásánál.

Ehhez hasonló eszközt alkalmazunk ma is, ha elemzőre van szükség, és az ismeretlen szavakat is megkíséreljük nyelvi információval ellátni. Például fordítórendszerekben természetes, hogy olyan tulajdonnevek szerepelnek, melyek toldalékolásánál nem találhatjuk a szótári szót a szótárunkban, de elemezni, generáláskor ragozni kell. A szóvégekből viszont jól lehet saccolni – csupán formai meggondolásokra hagyatkozva – melyik osztályba sorolhatjuk a szót.

4.12.2 Fák és ciklusmentes gráfok alkalmazása

Ha egyszerű nyelvi modellünk megengedi az asszociációs leírást, akkor az elemzés, generálás kulcsa a szóalakok, illetve elemzések tárolásának módjában rejlik. Sok alkalmazásban a szükséges szófát úgy építik, hogy minden elemhez automatikusan hozzárendelődik az ábécében elfoglalt helyének sorszáma. Ráadásul sorszám alapján könnyen kiválasztható a megfelelő tárolt kulcsszó. Így a két fában tárolt halmaz eleminek összerendelése egyszerűen elvégezhető. Akkor is alkalmas a tárolásra a fa, ha nem kell sok szót tárolni. Ilyen esetek, ha felhasználói, kiegészítő szótárat építünk munka közben, melybe felvehetjük az adatbázisunkban még nem szereplő szavakat, vagy egy szövegből kigyűjtött szavak elemzésére vagyunk csak kíváncsiak, és az esetleg bonyolultabb algoritmussal megelemzések eredményeit letároljuk, hogy a legközelebbi előfordulásnál ne kelljen annyit dolgozni, tehát „gyorsító tárba" mentjük le az előforduló elemzéseket.

Ha nem elemzésről vagy generálásról, csupán ellenőrzésről van szó, akkor gyakorlatban is felsorolható szóalakok esetén érdemes minimálautomatát építeni. Véges nyelv automatája ciklusmentes. Ha a ciklusmentes automata determinisztikus, akkor minimálása egyszerű feladat. Az általános algoritmus az állapotokat aszerint csoportosítja, hogy az állapot megkülönböztethető-e legfeljebb n hosszúságú sztringgel. Akkor különböztethető meg, ha létezik olyan n-nél nem hosszabb sztring, amelyik az egyik állapotból elfogadóba viszi az automatát, a másikból viszont elutasítóba. Nyilván ha $n=0$, akkor az elfogadó és elutasító állapotok oszlanak csak két csoportba. Az algoritmus dinamikus jellege abból áll, hogy ha megvan n-re az osztályozás, akkor $n+1$-re könnyen számítható a további osztályozás. Ráadásul könnyen belátható, hogy véges lépésben véget ér az algoritmus.

Az igaz, hogyha két állapot mindegyike elfogadó, vagy elutasító, ráadásul azonos karakterek olvasásakor egy lépésben pontosan ugyanabba az állapotba viszi az átmeneti függvény az automatát mindkét állapotból, akkor a két állapot (a minimálás szempontjából is) ekvivalens. Ez nem elégséges, csak szükséges feltétel a minimáltsághoz. Véges nyelv esetén viszont elégséges is, hogy ne legyen ilyen értelemben ekvivalens állapot. Ráadásul két állapot ilyen értelembe vett ekvivalenciájának megállapítása egyszerűbb, mint az általános minimálási algoritmus vizsgálata. Vagyis ciklusmentes determinisztikus véges állapotú automata minimálására elegendő az ilyen értelembe vett 1-ekvivalens állapotokat összevonni. Ezt is dinamikus programozás módszerével kell, mert ha két állapot egyesítünk, akkor ezáltal keletkezhetnek újabb 1-ekvivalens állapotok. Az algoritmus minden lépése csökkenti az automata állapotainak számát, ezért véges lépésben eredményre vezet.

A **Frey Tamás**tól, a Műszaki Egyetem Informatikai Tanszékének hajdani tanszékvezetőjétől származik a következő **algoritmus**, mely véges nyelvek esetén gyorsan képes előállítani azt a minimálautomatát (minimális ciklusmentes betűgráfot) mely a nyelvet reprezentálja. Azóta máshol is hallottam ezt a módszert, de én tőle tanultam.

Nevezzük szónak a nyelv elemeit. Feltételezzük, hogy a szavak betűrendes sorrendbe vannak rendezve, és nincs ismételt szó. (Ennél kevesebb is elég, mégpedig az, hogy az azonos kezdettel rendelkező szavak egymás után legyenek felsorolva.) A sorba rendezésre számos módszert találhatunk az irodalomban. Tulajdonképpen a betűfába (trie) rendezés is jó lenne, sőt maga a betűfa is egy véges automata, de nem minimális. Kis szókészletnél nem érdemes törődni a méretcsökkentéssel, a fa is megteszi, de nagyobb szólista esetén lényeges méretbeli különbséget érünk el a minimálással.

Az algoritmus során kétféle objektummal operálunk: véglegesített, illetve nem **véglegesített állapot**okkal. A végleges állapotokat akkor hozzuk létre, ha biztosak vagyunk abban, hogy ilyen állapotra szükség van, másrészt vele ekvivalens állapotot még nem hoztunk létre. Ezeket a nem végleges állapotok véglegesítésével tesszük meg. **Nem végleges állapot**okat használunk az a végleges automatába még be nem csatolt prefixek leírására, tehát ha adott prefixszel még nem vizsgáltuk meg az összes szóba jövő szót. Emiatt az ábécésorrendben érkező szavak esetén nem végleges állapotra csak annyira van szükség, amilyen hosszú az aktuálisan vizsgált szó. Ez egyben azt is jelenti, hogy nem véglegesített állapotból az algoritmus során legfeljebb annyira van szükség, amilyen a leghosszabb tárolandó szó hossza. Ezeket egy tömbben tároljuk, melyeket a prefix hosszával indexeljük.

Az algoritmus: Egy állapot azt az információt tartalmazza, hogy milyen karakterre melyik (véglegesített) állapotba megy át (tehát az átmenet parciális, de determinisztikus függvényt), valamint azt, hogy elfogadó állapot-e. A nem véglegesítettnél a legmagasabb indexűt kivéve az éppen beolvasott szó által meghatározott karakter viszi az eggyel nagyobb indexű nem véglegesítettbe. A végleges állapotokat egy tömbben tárolom.

Ha a nyelv nem üres, akkor mindenképpen szükség van egy olyan állapotra, amely elfogadó, de belőle nem lehet tovább lépni, és a teljesen specifikáltság miatt egy csapdára. Az általános elfogadó állapot lehet az első (nulladik) a véglegesített állapotok tömbjében, míg a csapda akár egy fiktív is lehet. Ezek tartalma triviális.

```
TemporaryStates[i] a nem véglegesített állapotok vektora.
States[i]     a véglegesített állapotok vektora.
MS            a véglegesített állapotok száma.
ACTWORD       az aktuális szó.
NEXTWORD      a következ• szó
ACTLENGTH     az aktuális szó hossza
NEXTENGTH     az következ• szó hossza
COMMONLENGTH  a két szó közös részének hossza
ACTSATE       az aktuális véglegesített állapot indexe
```

A két előre definiált állapot indexe: a `TRAP=-1`, és a `FIN=0`.

Egy állapot inicializálása nem más, mint annak bejegyzése, hogy egy karakterre sem lép tovább (minden karakterre a csapdaállapotra lép) és nem elfogadó állapot.

```
Subrutin CLEAR(State)
   FOR all letter State[letter]•TRAP; State.Accept•FALSE
```

Egy állapotot véglegesít, pontosabban, ha már létezik megfelelő állapot, akkor kiválasztja, ha nem, akkor létrehoz egy újat. Itt használjuk ki a korábbi elméleti meggondolást, elég a közvetlen átmenetek azonosságát ellenőrizni.

```
FUNCTION FIX(State)
      FOR I•0 (I<MS) & (States[I]•state) I•I+1;
      IF I=MS
      BEGIN
        States[MS]•State; MS•MS+1
      END
      RETURN I
```

1. Hozzuk létre az üres sztringet elfogadó állapotot, inicializáljuk az ideiglenes állapotokat, és olvassuk be az első szót. Ez az inicializálás:

```
      CLEAR(States[0]); States[0].Accept•True;
      ACTSTATE•0;
      MS•1;
      FOR I•0 TO MAXI CLEAR(TemporaryStates[I])
      ACTHOSSZ•0;
      ACTINDEX•0;
      COMMONLENGTH•0;
      READ(NEXTWORD)
      NEXTHOSSZ•Length(NEXTWORD);
```

2. Itt kezdődik a főciklus. A következő szót átírja az aktuálisba, és beolvassa a következőt. Ha nincs következő, akkor helyette az üres sztringet írja be. Jelzi a két szó közös prefixének hosszát, és az új szó hosszát.

```
  REPEATE
      ACTWORD•NEXTWORD; ACTHOSSZ•NEXTHOSSZ;
      READ(NEXTWORD); NEXTHOSSZ•Length(ACTWORD);
      FOR COMMONLENGTH•0 (COMMONLENGTH<ACTHOSSZ) &
                              (ACTWORD[COMMONLENGTH]=NEXTWORD[COMMONLENGTH])
                COMMONLENGTH•COMMONLENGTH+1;
```

3. Véglegesítsük a véglegesíthető állapotokat, vagyis az előző szó végétől, illetve attól az indextől, ahonnan nem volt véglegesítve, addig, amíg el nem érünk a közös prefixig.

```
      TemporaryStates[ACTHOSSZ].Accept=TRUE;
      FOR I•ACTHOSSZ STEP -1 I>COMMONLENGTH
      BEGIN
        ACTSTATE•FIX(TemporaryStates[ACTHOSSZ-1]);
        CLEAR(TemporaryStates[ACTHOSSZ]);
        TemporaryStates[ACTHOSSZ-1][ACTWORD[ACTHOSSZ-1]]•ACTSTATE;
      END
```

4. Ha volt még mit olvasni (a `NEXTWORD` nem üres) akkor ugorj a 2. pontra

```
  UNTIL NEXTHOSSZ • 0
```

5. Az így kapott állapot a kezdőállapota az automatának

```
  STARTSTATE•ACTSTATE;
```

Szemmel láthatóan a program a beolvasandó szavak összhosszával időben lineáris lenne, ha a véglegesítési `FIX(State)` rutinban nem kéne átvizsgálni, volt-e már megfelelő állapot a véglegesítettek közt. Ha tudjuk, hogy a bemenet teljes hossza L és az állapotok száma N, akkor az algoritmus $N \times Log(L)$-re csökkenthető. Lényeges azt is tudni, hogy a szótár méretétől függően egy nem erősen toldalékoló nyelvhez készített automata állapotainak száma 10 000-40 000 között szokott lenni.

A kritikus keresés persze gyorsítható. Az egyik lehetőség, ha a `TemporaryStates` táblázatban azt is jelöljük, vajon vezet-e ki olyan él, amelyhez tartozó átmenetnél nem találtunk a keresés során létező állapotot a véglegesítettek közt. Ebben az esetben ugyanis az aktuális prefixekhez rendelt állapot is egyedi kell, hogy

legyen, vagyis nem fogunk találni olyan állapotot a véglegesítettek közt, amely megegyezik vele. Így sok esetben megspórolhatjuk a keresés fáradságos műveletét. A keresést azzal is gyorsíthatjuk, ha az állapotokat nem egy, hanem több tömbben tároljuk aszerinti osztályokban, hogy hányfelé ágazik el az adott állapot, mint gráfpont. A keresés persze nem nagyságrendekkel, de lényeges konstanssal csökken. Ez a lépés azért is hasznos lehet, mert a különböző típusú állapotokat más-más módon érdemes reprezentálni, hogy kevés helyet foglaljon el. A tapasztalat szerint az élő nyelvek szókészleteiből készített automata állapotainak fele olyan, hogy egy irányba léphetünk tovább, vagyis csak egy karaktert fogad el az átmeneti függvény. Általában a kisebb fokszámú pontok sokkal többen vannak, mint a sokfelé ágazók, ezért ezekre lehet optimálni az implementációt.

4.12.3 Véges automata, mint helyesírás-ellenőrző

Rekurzív elemet tartalmazó konkatenációs modellnél remény sincs DAG használatáról. Az említett formalizmusok viszont tálcán kínálják a lehetőséget. Ahogy a 4.10.14 fejezetben említettem, véges automatás reprezentáció lehetséges megoldása lehet a helyesírás-ellenőrzőnek. Ha bármely formalizmust (HUNSPELL, HUMOR) átkonvertáljuk egy olyan leírásra, ahol folytatási osztályok vannak, és azok között az átmenet véges nyelven megadható, akkor a következő megadási formát kapjuk:

$$A_i=\beta_i+\sum_j A_j\alpha_{ij} \text{ ahol a}$$

$\sum$ és + uniót jelent
A_i kategóriát (résznyelvek osztályát)
α_{ij}, β_i jól definiált nyelvet (jelen esetben olyan morfémahalmazt, mely egy szót az egyik kategóriából a másikba viszi, illetve a kategóriához tartozó szótövek halmazát)
az egymás után írást a nyelvek konkatenációját.

Az ilyen séma hasonló a véges automatáknál megszokott leírásához, azzal a különbséggel, hogy az állapotok közötti átmenet nem karakter(halmaz) alapján, hanem az α_{ij} nyelv egy elemének „beolvasásával" történik.

Ha az egyenletrendszer baloldalán szereplő A_i kategóriák csak egy egyenlet baloldalán szerepelnek, az ilyen reguláris egyenletrendszernek a nyelvek halmazában mindig van megoldása. Vagyis vannak olyan A_i nyelvek, melyek kielégítik a felírt egyenletrendszert. A séma **rekurziómentes**, ha belőle nem származtatható behelyettesítésekkel olyan $A_i=\beta_i+\sum_j A_j\alpha_{ij}$ egyenlet, melynek jobb oldalán az α_{ii} nem üres nyelv, tehát van eleme. A séma **ε-ciklustól mentes**, ha nincs olyan A_i kategória, amelyből elindulva egy vagy több lépcsőben úgy jutunk el ismét A_i-be, hogy az átmeneti α nyelvek mindegyike tartalmazza az ε-t. Nos, ha a séma ε-ciklustól mentes, akkor egyértelmű megoldása van a felírt egyenletrendszernek, ha az egyenletek száma és a kategóriák száma megegyezik. Ha ráadásul egy rekurziómentes sémában az α-k, β-k véges nyelvek, akkor az egyenletrendszernek is véges nyelv a megoldása.

Nos, a konkatenációs modellekre igaz, hogy az α_{ij}-k végesek, így reguláris nyelvek. Sőt, ha van ciklus, az csak képzősorozatból állhat, amelyek nem lehetnek nullahosszúak. Ebben az esetben az is igaz, hogy az osztályok is reguláris nyelvek, és az egyenletnek egyértelmű a megoldása.

Az Elekfi-féle paradigmatikus osztályozásból származó szóalakokra vonatkozó egyenletrendszerben 400 alatti kategória van, melyeket véges nyelvek kötnek össze. Ez magában hordozza a véges állapotú automata implementációt. Nevezzük nyelvi állapotnak a paradigmából származó osztályokat. Innen egy véges nyelvvel lehet másik, esetleg ugyanabba az osztályba eljutni. A véges nyelvnek megfelelő részautomatát akár a Frey módszerrel is generálhatjuk, azzal a különbséggel, hogy a nyelv elemei után megjelöljük, melyik új nyelvi osztályba jutunk a beolvasás után. Így maga a beolvasásnak ez a szakasza determinisztikus, ha tudnánk, hol a morféma vége a sztringben, illetve ha beolvastuk az osztályból kiinduló nyelv egy elemét, egyértelmű lenne, melyik osztályba jutunk. sajnos mindkét eset miatt lehet indeterminisztikusság. Az indeterminisztikusság ilyen modellel nem is kerülhető el, mert a szóazonosítás követi a szó nyelvi elemzését. Márpedig ha egy szóalaknak több elemzése van, akkor érthető, az ilyen modell indeterminisztikus.

Az indeterminisztikusság miatt az automata elvileg több szálon paralel bejárható utak felderítésével történhet, de ma még nincs olyan sokprocesszoros gép, amellyel érdemes lenne ezt megtenni. Helyette egyszerű visszalépéses (backtrack) módszert alkalmaznak. Azért lehet gyorsítani. Ellenőrzés esetén a következő két módszer vált valódi gyorsítóvá:

1. **Válaszd először a hosszabbat** – vagyis amikor egy kategóriának keressük a folytatását, először a hosszabb morfémát választjuk, a rövidebbek folytatási lehetőségét a backtrack-re hagyjuk. Ez a szótövek keresésénél a leghatékonyabb, de máshol is segít.

2. Az elemzésnél **válaszd a valószínűbb utat** – vagyis ha egy morfémaalak több osztályba vezet, akkor a gyakoribbat válaszd, a másikat csak sikertelenség esetén próbáld ki.

Ha szóellenőrzésre használjuk az így optimált automatát, a tapasztalat szerint 20 százalékkal lassabb a determinisztikus változatnál, viszont memóriaigénye több nagyságrenddel kisebb. Ha a magyar nyelv szavait klasszikus minimálautomatával szeretnénk felismerni, az állapotok száma meghaladná a 100 000 000 000-ot. Így viszont 100 000 körüli az érték. A sebességi enyhe romlás persze csak ellenőrzés esetén érhető el. Elemzéskor minden utat be kell járni. Hasonló az eset, ha ellenőrzéskor ismeretlen szó kerül vizsgálatra. Normál esetben viszont a szövegben található ismeretlen szavak százalékos aránya elenyésző az helyesekhez képest.

A nyelvi leírásokban nem csak jobb oldali, hanem bal oldali bővülés is lehetséges. Ha a szóösszetételt nem számítom, a prefixek – magyarban az igekötők, mellékneveknél a felsőfok kezelés esetére kell gondolni.

Ha a sémában megengedjük a $A_i=\gamma_{ij}A_j$ egyenletet, de nem származik ilyen szabályokból rekurzió a korábbi meghatározás szerint, akkor ez mit sem változtat a korábbi állításokon. Tehát sem azon, hogy van-e megoldása az egyenletrendszernek, sem azon hogy mikor egyértelmű a megoldás, és ha a γ-k is reguláris nyelvek, akkor az egész rendszernek reguláris a megoldása.

4.12.4 Állapotcsökkentő módszerek

Az előző fejezetben említettem, hogy klasszikus determinisztikus automata konstrukciója magyar nyelv esetén technikailag lehetetlen. Legfeljebb akkor használható, ha a nyelvet erősen megszorítjuk, korlátozzuk. Egyéb esetben az állapotok száma kezelhetetlenül nagy lenne. Nem determinisztikus automata persze lehet elfogadható méretű. Ha arra teszünk kísérletet, hogy mely pontokban váljék indeterminisztikussá az automatánk, tehát hol ne törődjünk azzal, hogy nem determinisztikus, akkor a gyakorlatban az vált be, hogy a morfotaktika által meghatározott pontok, tehát a morfémahatárok a legalkalmasabbak erre.

Más módszerek is vannak, mely az állapotok csökkentéséhez vezetnek.

Véges nyelveknél kísérleteznek azzal, hogy kevesebb állapotú olyan véges automatát állítanak elő, mely a véges nyelv minden elemét elfogadja, de olyan szót is, melynek hossza meghaladja a véges nyelv elemeinek hosszát, de nincs benne a nyelvben. Egy véges nyelv **lefedő véges állapotú automatá**ja olyan determinisztikus véges automata, amely pont ezt a feltétel tejesíti. Tehát ha egy sztring nem hosszabb, mint a véges nyelv szóhosszainak maximuma, akkor éppen akkor fogadja el a szót, ha az a nyelv eleme. A szóhossz könnyen ellenőrizhető, így ebből egy szóellenőrző könnyen készíthető az eredeti nyelvre. Ráadásul a lefedő automaták között is lehet definiálni **minimális lefedő automatá**t. Ez nem unikális, több is létezhet, de aránylag egyszerű algoritmussal megkonstruálható valamelyik.

A módszer általánosítható. A karakterszámláló felfogható egy véges automatának. Ebben az értelemben a fenti módszer abból áll, hogy egy automata helyett kettőt konstruálunk, és akkor fogadjuk el a szót, ha mindkét automata elfogadja. Ha a két automata determinisztikus, akkor az egész algoritmus is az. Más szóval, ha létezik egy A automata a hozzátartozó L_A nyelvvel, és egy B automata, amelyhez tarozó L_B nyelv $L_B \geq L_A$, akkor kereshetünk olyan (minimális állapotú) C automatát, hogy a hozzátartozó L_C olyan, hogy $L_A=L_B \cap L_C$. Vagyis keresünk az A automatához B automata megszorításával egy C lefedő (minimál)automatát. Esetleg a B sem ismert, tehát egy véges automatát fel szeretnénk bontani két automata metszetére. Mint közismert, két (minimál)automata nyelvének metszetét elfogadó (minimál)automata állapotainak száma a két automata állapotszámának szorzatával felülről becsülhető, de sok esetben ez alá nem is mehetünk. Ha egy sokállapotú automatát sikerül **faktorizálni**, tehát kisebb automaták szorzatára bontani, akkor az összállapotszám nagyságrendekkel csökkenthető. Az elemzési algoritmus valamivel bonyolódik, időben nagyjából annyival szorzódik, ahány paralel automatát használunk, de helyigénye ennél nagyobb nyereség. Az automaták felbontására viszont nem láttam elviselhető sebességű algoritmust. A feladat algoritmikusan megoldható, de általában **NP-nehéz probléma**.

A felbontást ellentétes irányból viszont meg lehet közelíteni. Nem szétbontjuk az automatát, hanem eleve paralel algoritmusokat gyártunk, aminek eredményeként mégis megkapjuk a kívánt elemzőt.

Ezzel a trükkel lehet megoldani a prefixek kezelését is. Ha például a szó elején találunk egy *legesleg*-et, akkor egy mellékautomata átmegy egy nem elfogadó állapotba, és akkor megy elfogadóba, ha elolvassa a középfok *bb*-jét. A „főautomata" ettől függetlenül működik. Ha ezt a faktorizálást nem tennénk meg, az állapotok számát meg kéne háromszorozni. Hasonló harmadoló eredménye van az igekötők leválasztásának is.

Másik lehetőség az **automata többszintű**sége. Pontosabban a véges automatának rokona a véges fordító. Mivel reguláris nyelv véges fordítása is reguláris, megtehetjük, hogy az inputot lefordítjuk egy másik reguláris nyelvre, és azon végezzük el az elemzést. Ez a módszer nagyon hasonlít a fordítóprogramok működéséhez, ahol a forrásnyelvről először egy közbülső nyelvre fordítanak, majd onnan a célnyelvre. Ez a módszer nagyon rugalmas, és sok nyelvi jelenség kezelésére használható.

ris nyelvre, és a lefordított nyelvet véges állapotú elemzőnek vetjük alá. Az ilyen értelembe vett több szintre való bontásnak sincs általános gyors algoritmusa, hisz a probléma megoldásának nehézségi foka megegyezik a metszet faktorizálásával, csupán az ismert, hogy a két szint állapotainak szorzatával lehet felülről becsülni az egyesített mininimálautomata állapotainak számát. Maga a fordító is elutasíthatja az olvasott szót, de ha átmegy rajta, akkor a második szint dönt, helyes-e a szó. Itt is inkább a direkt konstrukció a bevált módszer. A gyakorlatban az első szint a lokálisan ellenőrizhető dolgokban dönt, míg a másodikra bízzák a távoli egyeztetéseket. Ezzel a módszerrel a magyarban a második szintre szokták tenni a prefixek és a helyes szóösszetételek ellenőrzését. Elvileg az is megoldható, hogy a fonetikai és szótári ellenőrzéseket az első, a fordító szint végzi, míg a morfotaktikai megfelelés a második szint feladat, amely az első szinttől fordításként a morfémák nyelvtani kategóriáit kapja meg bemenetként.

4.12.5 Kétszintű morfológia (TLFSA)

A 80-as végén kezdett recsegni a listás módszer uralma. Sok helyen kísérleteztek különböző eszközzel túllépni az amerikai hozzáálláson, azon, hogy mindent fel lehet sorolni, ezért nincs szükség morfológiai eszközökre. Kimmo Koskenniemmi és Lauri Kartunnen a fonetikusok hozzáállásából indult ki. A leírt karakter és a kiejtett hang közötti meghatározottság a legtöbb nyelven a közvetlen szomszédos betűktől, hangoktól függ. Emiatt az írás és a kiejtés két szintjét a szűk környezet alapján kell megadni.

A morfológia esetén, ahol persze nincs szó kiejtésről, megkülönbözteti a morfémák szótári (mély) alakját a felszíni alaktól. E két szint szinkronját próbálja leírni, és nem azt, hogyan alakul az egyikből a másik. Ha elemezni kell, akkor a felszíni alakhoz keresi azt a szótári alakot, amely a leírás szerint megfelel a felszíni alaknak és a környezetnek egyaránt. Generáláskor a szótári alakhoz keresi a megfelelő felszíni alakot. A két alak közt elsősorban hangtani kapcsolat van, melyet a környezet meghatároz.

A formalizmust egy magyar példán keresztül mutatom be. A *-ra*/*-re* helyrag szótári alakját jelöljük úgy, hogy nincs meghatározva, a felszíni alak *ra* vagy *re*; a szótári alak legyen *ræ*. Ha magas hangrendű szóhoz kapcsolódik, akkor *ra*, ha mélyhez, akkor *re*.

```
æ:a<=>BC*r_
```

A leírásban a szabályoknak van bal és jobb oldala. A kettőt egy- vagy kétirányú nyíl választja el. A baloldali rész a szótári karakter és felszíni karakter megfelelését jelzi. A jobb oldal a megfelelés elégséges **<=**, szükséges **=>** vagy mindkét minőségét **<=>** jelzi az elválasztó nyílnak megfelelően. A feltételnek a centrumát az aláhúzás _ mutatja. Ez az a pont, ahol a szótári (mély) karaktert megfeleltetjük a felszínivel a szabály bal oldalának megfelelően. A centrum előtt és után olyan reguláris kifejezések szerepelnek, melyekben a szótári és felszíni karakterek párjaiból építkezik, némi gyorsírással. A gyorsírás azt jelenti, hogy bizonyos karakterhalmazokat (általában nagybetűvel jelölve) tömörebben jelezhetünk. A példasor azt mutatja, hogy az `æ` szótári karakter felszíni megfelelője az `a` felszíni karakter, akkor és csak akkor, ha őt egy `r` előzi meg, azt egy morfémahatár – melynek felszíni alakja az üres sztring, azt valahány, akár 0 db mássalhangzó, és azt egy mélyhangrendű magánhangzó. A leíráshoz még annyit, hogy ha nem használjuk a `:`-ot, az azt jelenti, hogy a szótári és felszíni karakter azonos. tehát a `x` önmagában ugyanaz, mintha `x:x`-et írnánk. A másik megjegyzés, hogy két deklarált karaktere van a formalizmusnak. Az egyik a **@**, a másik az **ε**. A **@** jelentése a bármi, az **ε** pedig a semmi (kihagyás, törlés). A morfémahatár, ha jelöljük, elnyelődik a felszínen. Ezt jelenti a **|:ε**. A fenti szabály nem pontos, mert nem a rag előtti utolsó magánhangzó határozza meg a szó hangrendjét, de a példa egyszerűsége miatt tekintsünk el ettől a hibától. Ennek mintájára egy másik szabályt is felírhatunk:

```
æ:a<=>BC*|:εb_
```

ami a *-ba*/*-be* és *-ban*/*-ben*-t intézi el, vagy összevonva a korábbival, ami még sok más toldaléknak is jó:

```
æ:a<=>BC*|:εC_
```

Ilyen módon a szótárban csak `ræ, bæ, bæn, næk, væl` kell, a felszíni megfelelő egy szabállyal választódik ki. Hasonlóan oldható meg a *ról*, *tól* stb. A tőváltozások egy része is leírható ezzel a formalizmussal:

```
v:ε<=>bb|:ε_      v:c=>c|:ε_      v:ε<=>ff|:ε_
 v:b=>b|:ε_      v:ε<=>dd|:ε_      v:f=>f|:ε_
v:ε<=>cc|:ε_      v:d=>d|:ε_            …
```

azt adja meg, hogy a morfémahatárt követő `v`-ből teljes hasonulás lesz, ha szimpla mássalhangzóra végződik a szó, és elnyelődik, ha dupla mássalhangzó előzi meg a morfémát.

A magánhangzóra és mássalhangzóra szabályok leírása független, de a példánkban is független a hangrendi illeszkedés és a mássalhangzó-hasonulás. A módszer egyik nagy előnye éppen ebben rejlik. Ha sikerül felszíni morfémaalternánsok meghatározását kisebb ortogonális tulajdonságok alapján kiválasztani, akkor paralel kisméretű automatákkal tudjuk elvégezni az ellenőrzést ahelyett, hogy egyetlen nagy automatát kezelnénk. Ennek akkor van értelme, ha egy szabályt valóban egyszerű automatával lehet implementálni.

Az automatát viszont meg is alkották. Az automata nem a szokásos. Nem egy fordító, amely a szótári (mély) leírásból generálja a felszíni alakot, sem fordítva, amely elemzéssel a felszíni alakból generálja a szótári alakok szekvenciáját. Az automata a megfeleltetést reprezentálja a felszín és a szótári alakok között.

Ha elméletileg akarom megközelíteni, akkor azt mondhatom, hogy egy olyan automatát rendelünk a szabályokhoz, amely a **fordítást jellemző nyelv**et fogadja el. A fordítást jellemző nyelv egy olyan nyelv, amely szavaiból a fordítandó és fordított nyelv egy nagyon egyszerű véges fordítóval állítja elő. Az egyszerű véges fordítónak nincsenek különböző állapotai, tehát minden lépésben ugyanúgy viselkedik azonos karakterre. Egy karakter beolvasásnál vagy egy karaktert ad ki, vagy ε-t, üres sztringet. Vagyis a fordítást jellemző nyelv olyan nyelv, mely szavaiból egy-egy egyszerű homomorfizmussal kinyerhetjük a forrás és a célnyelv egymáshoz tartozó elemeit. Megjegyzendő, hogy ebben az értelemben könnyen felcserélhető a forrás és a célnyelv, nincs igazán megkülönböztetve.

Jelen esetben a fordítást jellemző nyelv ábécéje a két ábécé direkt szorzata – pontosabban bővítve azzal, hogy lehet az egyik vagy a másik oldal ε is –, melyet jelölésünkben a karakterpárt `:`-tal elválasztva írjuk le. A jellemző nyelv szavaiból úgy állítom elő a két nyelv egymásnak megfelelő elemét, hogy az egyik esetben a direktszorzat egyik tényezőjét, a másik esetben a másik összetevőjét adom ki a karaktereknek, ε esetén pedig semmit. A kétszintű morfológia szabályai könnyen átírhatók ezen dupla karakterek feletti reguláris kifejezésekké, még azt is belevéve, hogy nem a teljes szóra, hanem csak egy részletére vonatkoznak. Tehát minden szabály megfogalmazható reguláris kifejezésként. Ha véges sok szabály van, akkor véges sok reguláris nyelv metszetét jelenti a szabályok követése. Ez pedig reguláris nyelvet határoz meg. Még akkor is, ha az egyes morfémák szótári alakjait rögzítem, mert ezek véges nyelvek. Ha a fordítást jellemző nyelvből a korábban említett véges fordító állítja elő a forrás és célnyelv elemeit, tehát ezek is regulárisak, sőt elvileg az egyikből a másikat véges fordítóval elő tudom állítani.

Az alapvető feladat, hogy a gyors végrehajtás érdekében a szabályokhoz elkészítsük a véges automatát. A formalizmus kitalálásakor még gyengébb volt egy PC ereje. Az a fordító, amely előállította a formális leírásból az automatákat, csak nagyobb szervereken futott. Ezért a rendszernek egy olyan változatát hozták létre, melyben az automatát előállítani ugyan nem lehetett, de futtatni már igen, Ez a PC-KIMMO. Ebben arra is van lehetőség, hogy kézzel szerkesztett automatát adjunk meg, és azt futtassuk. Ezen automaták olvasható alakja táblázatos. A sorok az állapotokat, az oszlopok a karakterpárokat jelentik. Persze a pároknál itt is szerepelhetnek a fent említett tömörítések, vagyis karaktercsoportokat is használhatunk, viszont emiatt lehet, hogy egyes konkrét karakterpárok többször is szerepelnek. Ilyenkor a karakterpárnak megfelelő első megtalált oszlop az érvényes. A mátrix elemei azt jelölik, melyik állapotba megy át az automata. és egy külön oszlopban ponttal jelöljük, hogy a részautomata egy állapota elfogadó-e. A formalizmushoz némi deklaráció is jár – a nyelv abc-je, a speciális karakterek és betűhalmazok definiálása. Egy részlet a magyar morfológiai leírásból:

```
ALPHABET
  a á e é o ó ö ő u ú ü ű i í t l s r z n k g m d v b j f y c p h x w q ä
  A Á E É O Ó Ö Ő U Ú Ü Ű I Í T L S R Z N K G M D V B J F Y C P H X W Q Ä
  0 1 2 3 4 5 6 7 8 9 . ' - ` & | + Ę ş § { } Ľ ß Ť Ř ř Ń ń
; {Le*}Le*+& the letters in parantheses appear on surface if only if +&
;' = apostrophe
; +` lengthens the last a, e, o ö - if one of them is at the end
; +Ę mark for nomilal consonant assimilation (before val,vel,vá,vé)
; +ş stends for verbal consonant assimilation of j in imperative or obj.infl
; +§ stends for verbal consonant assimilation of j in imperative special case
; + stands for affix start
; | stands for compound boundary
; Ř stands for noun
; Ľ stands for adjactiv
; ß stands for comperativ
; Ť stands for verb
; ř stands for number
; Ń stands for verbal prefix
; ń stands for prefix of superlativus
NULL ^ ; the digits may be constituants of words!
ANY @
BOUNDARY #
SUBSET Co t l s r z n k g m d v b j f y c p h x w q T L S R Z N K G M D V B J F Y C P H X W Q
SUBSET Co0 t l s r z n k g m d v b j f c p h x w
```

```
SUBSET Co1 t l s z n g c
SUBSET Co2 s z y h
SUBSET Cd1 r k m d v b j f c p h w
SUBSET Vo a á e é o ó ö ő u ú ü ű i í ä A Á E É O Ó Ö Ő U Ú Ü Ű I Í Ä
SUBSET Vc ş §
SUBSET Nc ` Ę
SUBSET CL Ľ ß ľ Ř ř Ń ń
RULE "Default" 1 102
   a á e é o ó ö ő u ú ü ű i í t l … x w q - A Á E É O Ó Ö Ő U Ú Ü Ű I Í T L … X W Q ' 0 1 … 9 . ä Ä ` & | + { } Ę ş § Ľ ß ľ Ř ř Ń ń @
   a á e é o ó ö ő u ú ü ű i í t l … x w q - A Á E É O Ó Ö Ő U Ú Ü Ű I Í T L … X W Q ' 0 1 … 9 . ä Ä ` & ^ ^ ^ ^ v ^ ^ ^ ^ ^ ^ ^ ^ @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 … 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 … 1 1 1 1 1 1 … 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
; The optional part of a morpheme!!!
; &:^ <= {:^ Le* }:^ Le* +:^ _
; Le:^ <= {:^ (Le:^)* _ (Le:^)* }:^ Le* +:^
; &:^\ <= Le:^ }:^ Le* +:^ _
; Le:^ => {:^ (Le:^)* _ (Le:^)* }:^ Le* +:^
&:^
RULE "Insert" 4 10
   +  {  }  `  & Le Le  CL @
   ^  ^  ^  ^  &  ^ Le  ^  @
1: 1  2  0  1  1  1  1  1  1
2. 0  0  1  0  0  3  6  2  0
3: 0  0  4  0  0  3  0  3  0
4: 5  3  0  4  0  4  4  4  4
5. 1  2  0  1  0  1  1  5  1
6. 0  0  7  0  0  0  6  6  6
7. 8  6  0  7  0  7  7  7  7
8. 0  0  0  0  1  0  0  8  0
```

```
; a-á e-é o-ó ö-ő
; a:á => _ +:^ `:0    a \ <= _ +:^ `:0
; e:é => _ +:^ `:0    e \ <= _ +:^ `:0
; o:ó => _ +:^ `:0    o \ <= _ +:^ `:0
; ö:ő => _ +:^ `:0    ö \ <= _ +:^ `:0
RULE "Long1" 5 13
   a  a  e  e  o  o  ö  ö  +  `  Ę CL  @
   a  á  e  é  o  ó  ö  ő  ^  ^  v  ^  @
1: 4  2  4  2  4  2  4  2  1  1  1  1  1
2. 0  0  0  0  0  0  0  0  3  0  0  2  0
3. 0  0  0  0  0  0  0  0  0  1  1  3  0
4: 4  2  4  2  4  2  4  2  5  0  0  4  1
5: 4  2  4  2  4  2  4  2  0  0  0  5  1
```

```
; s+j z+j megy+j jön+j
RULE "Jforbid" 8 27
    z s + z s z + s z + d t t ö m n e j y ú v + Ę Ę CL g @
    z z z ^ s s s ^ d d ^ s z ö m n e j y ú v j z ^  ^ g @
1:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 6 6 8 1 1 2 2  1 1 1
2:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 6 6 8 1 0 2 2  2 1 1
3:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 4 6 6 8 1 1 2 2  3 1 1
4:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 6 6 8 1 1 2 2  4 5 1
5:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 6 2 8 1 1 2 2  5 1 1
6:  2 2 2 2 2 2 2 2 2 2 2 2 2 7 3 1 1 6 6 8 1 1 2 2  6 1 1
7:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 1 6 6 8 1 1 2 2  7 1 1
8:  2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 6 6 8 2 1 2 2  8 1 1

;  Vo:^ <=> Co_Co*+:^? …eher-erh ehely-elyh ohor-orh oholy-olyh orom-orm
;RULE "VowelEllision" 16 24
     e  o  h  h  h  r  r  l  l  y  y  e  e  o  o  Co Vo Vo |  +  {  }  @
     y  y  h  r  l  r  h  l  h  y  ^  e  ^  o  ^  Co Vo ^  ^  ^  ^  ^  @
 1:  1  1  1  1  1  1  1  1  1  1  1  3  2 10  2  1  1  2  1  1  1  1  1
 2.  0  0  2  0  0  2  0  2  0  2  0  0  0  0  0  2  0  0  2  16 0  0  0
 3:  1  1 13  4  7  1  1  1  1  1  1  1  2  1  2  1  1  2  1  1  1  1  1
 4.  0  0  0  0  0  0  0  0  0  0  0  0  5  0  0  0  0  0  0  0  0  0  0
 5.  0  0  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 6.  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  16 0  0  0
 7.  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 8.  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 9.  0  0  0  0  0  0  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0
10:  0  0 15 11 12  1  0  1  0  1  0  1  0  1  0  1  1  0  1  1  1  1  1
11.  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  0  0  0  0  0  0  0  0
12.  0  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
13:  1  1  1  1  1  1  1  1  1  1  1  3 14 10  2  1  1  2  1  1  1  1  1
14:  0  0  2  0  0  0  0  0  0  2  0  0  0  0  0  2  0  0  14 16 0  0  1
15:  0  0  1  0  0  1  0  1  0  1  0  3  2 10 14  1  1  2  1  1  1  1  1
16.  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1

; The first letter may be capitalized
RULE "Capitalize" 2 37
   a á e é o ó ö ő u ú ü ű i í t l s r z n k g m d v b j f y c p h x w q ä @
   A Á E É O Ó Ö Ő U Ú Ü Ű I Í T L S R Z N K G M D V B J F Y C P H X W Q Ä @
1: 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2

; When a proper name has derivation to adjective or verb
RULE "Lower" 5 41
   A Á E É O Ó Ö Ő U Ú Ü Ű I Í T L S R Z N K G M D V B J F Y C P H X W Q Ä Ľ ß ľ Ř @
   a á e é o ó ö ő u ú ü ű i í t l s r z n k g m d v b j f y c p h x w q ä ^ ^ ^ ^ @
2: 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
2. 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2
3: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 3 3 3 3
4. 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 4 4
5: 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

; el|felejt
RULE "Verbal" 5 11
    Ľ  ß  ľ  Ř  ř  Ń  ń  -  |  |  @
    ^  ^  ^  ^  ^  ^  ^  -  -  ^  @
1:  4  4  5  4  4  2  4  4  4  4  1
2:  2  2  0  2  2  0  0  1  1  3  2
3.  3  3  5  3  3  0  0  3  3  3  3
4:  4  4  5  4  4  3  4  4  4  4  4
5:  4  4  5  4  4  3  0  4  4  0  5

; leg…bb
RULE "Super" 2 3
```

```
    ß  ń  @
    ^  ^  @
1:  1  2  1
2.  1  2  2

; sakk-kör, balett-tan…
RULE "Triplett" 11 7
    k  t  y  |  |  CL @
    k  t  y  ^  -  ^  @
 1: 2  5  1  1  1  1  1
 2: 3  5  1  1  1  2  1
 3: 3  5  1  4  1  3  1
```

```
 4: 0  5  1  1  1  4  1
 5: 2  6  1  1  1  5  1
 6: 2  6  9  7  1  6  1
 7: 2  8  1  1  1  7  1
```

```
 8: 0  0  1  0  0  8  0
 9: 2  5  1 10  1  9  1
10: 2 11  1  1  1 10  1
11: 2  5  0  1  1  1  1
```

Az automatáknál az első nem megjegyzéssor a mátrix méretét adja meg, a második a mélyreprezentációs karakter(halmaz)t, harmadik a felszíni karaktert, majd sorszámozva az átmeneti függvény mátrixának sorai. A sorszámot `:` követ, ha nem elfogadó állapot, `.` követ, ha elfogadó sorszáma. A deklarációnak megfelelően `@` a bármi, és `^` az üres jele. A példából néhány helyen látszik a szintaktikus megadása a szabálynak, de sok helyen egyszerűbb volt közvetlen az automatát megadni, mintsem a formalizmusban leírni a szabályt. Néhány speciális karaktert is használtunk az olyan nyelvi információk kezelésére, melyek a szóformából nem derülnek ki, ezért a szótári mélyreprezentációban szerepelnek.

A szabályokon kívül szükséges a szótár megadása is. Pontosabban Kimmóék is rájöttek, hogy a konkatenációs modellből adódó információ, a morfotaktika nem vihető be ebbe a modellbe közvetlenül, ezért különböző kategóriájú szótárak vannak, melyekben a tételek utalnak a folytatási osztályokra, résszótárakra. Példa a tőtárból:

```
\lf lappangľ
\alt I17
\lx ROOT
\gl V(lappang)

\lf lappangľ
\alt Ij97
\lx ROOT
\gl V(lappang)

\lf laptikaŘ
\alt XN6A
\lx ROOT
\gl N(laptika)
```

```
\lf lapuŘ
\alt n1A
\lx ROOT
\gl N(lapu)

\lf lapulľ
\alt i1
\lx ROOT
\gl V(lapul)

\lf lassítľ
\alt I38
\lx ROOT
\gl V(lassít)
```

Egy szótári tétel három részből áll:

1. `lf`=szótári alak. Ezt használja az automata az azonosításnál. Emiatt lehetnek benne kibővítő karakterek.
2. `alt`=folytatási osztály kategóriája. A szótári szinten, ha ezt az egységet azonosította, akkor az automata a jelzett szótárban keres tovább.
3. `lx`=a szótár kategóriája. Jelzi, hogy az adott morféma melyik szótárhoz tartozik
4. `gl`=olvasható forma. Az elemzési szinten ez a szép kiírási forma.

Az angol szabályrendszerre egy példa:

```
;english.rul
;Rules file for Englex
;last modified 28-Nov-95
;Englex 2.0b5
;Copyright (C) 1991-1995, Summer Institute of Linguistics, Inc.
;Evan Antworth                    |  e-mail: evan.antworth@sil.org
;Academic Computing Department    |  phone:  214-709-3346, -2418
;Summer Institute of LInguistics  |  fax:    214-709-3363
;7500 W. Camp Wisdom Road
;Dallas, TX 75236
;This description of English is based on the article "A two-level
;morphological analysis of English," by Lauri Karttunen and
;K. Wittenburg, Texas Linguistic Forum 22:217-228 (1983).
;See appendix A of the PC-KIMMO book for an exposition of the rules
;in this file (though this version differs).
;Added upper-case letters 5-Apr-95
;CONTENTS
; Defaults
; Epenthesis
; y:i-spelling
; Elision
; Gemination
; END
; s-deletion
; i:y-spelling

;Symbol key:
; ' = apostrophe
; - = hyphen
; ˋ = stress
; + = morpheme break
; . = period (used in abbreviations such as U.S.A.)

ALPHABET
 ;lexical (upper) and surface (lower) characters:
     b c d f g h j k l m n p q r s t v w x y z a e i o u ' - .
     sh ch  ;digraphs
```

```
    B C D F G H J K L M N P Q R S T V W X Y Z A E I O U
 ;lexical (upper) only characters:
      +
NULL     0
ANY      @
BOUNDARY #
SUBSET CN     b c d f g h j k l m n p q r s t v w x y z sh ch
SUBSET CNsib  s x z sh ch              ;sibilant consonants
SUBSET CNpal  c g                      ;palatal consonants
SUBSET CNgem  b d f g l m n p r s t    ;geminated consonants
SUBSET VO     a e i o u
SUBSET VObk   a o u                    ;back vowels

RULE
"Defaults 1" 1 33
    b c d f g h j k l m n p q r s t v w x y z sh ch a e i o u ' - ` + @
    b c d f g h j k l m n p q r s t v w x y z sh ch a e i o u ' - 0 0 @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
;uppercase letters and period
RULE
"Defaults 2" 1 28
    B C D F G H J K L M N P Q R S T V W X Y Z A E I O U . @
    B C D F G H J K L M N P Q R S T V W X Y Z A E I O U . @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

;This +:- rule is useful for Recognition, as it permits recognition of affixes
; with an optional hyphen, e.g. pre-pay or prepay.
;If you use this rules file for Generation or Synthesis, comment out this rule to
;avoid multiple results.
;RULE
;"+:-" 1 2
;    + @
;    - @
; 1: 1 1
;Special correspondences are declared in the column headers of the rules that use them.
;(The s:0 and i:y correspondences are found in the optional rules after the END keyword.)
;;   +  y  e  +  +  +  +  +  +  +  +  +  +  +  s  i
;;   e  i  0  b  d  f  g  l  m  n  p  r  s  t  0  y

;==========
;Epenthesis
;==========
; LR: fox+s  kiss+s  church+s  spy+s
; SR: foxes  kisses  churches  spies
RULE
"+:e <=> [CNsib | y:i | o] +:@ ___ s [+:@ | #]" 7 8
     +  CNsib  +    s    #    y    o   @
     e  CNsib  @    s    #    i    o   @
 1:  0    2    1    2    1    2    7   1
 2:  3    2    5    2    1    2    7   1
 3.  0    0    0    4    0    0    0   0
 4.  0    0    1    0    1    0    0   0
 5:  0    1    1    6    1    1    1   1
 6:  0    1    0    1    0    1    1   1
 7:  3    2    1    2    1    2    7   1


;============
;y:i-spelling
;============
; LR: spy+s  happy+ly  spot+y+ness
; SR: spies  happiOly  spottiOness
RULE
"y:i <= @:CN ___ +:@ ~[i | ']"  4 7
     @   y   y   +   i   '   @
     CN  i   @   @   i   '   @
 1:  2   1   1   1   1   1   1
 2:  2   1   3   2   1   1   1
 3:  2   1   1   4   1   1   1
 4:  0   0   0   0   1   1   0

;=======
;Elision
;=======
; LR: `move+ed `move+ing `move+able `be+ing
; SR: Omov0+ed Omov0Oing Omov0Oable 0beOing
; LR: `trace+ed `trace+ing `trace+able
; SR: OtracOOed OtracOOing OtraceOable
RULE
"e:0 <= VO CN CN* ___ +:@ VO" 8 8
      CN  CNpal e   e   +   VO VObk @
      CN  CNpal 0   @   @   VO VObk @
 1:   1    1    1   2   1   2   2   1
 2:   3    6    1   2   1   2   2   1
 3:   3    6    1   4   1   2   2   1
 4:   1    1    1   2   5   2   2   1
 5:   1    1    1   0   1   0   0   1
 6:   1    1    1   7   1   2   2   1
 7:   1    1    1   2   8   2   2   1
 8:   1    1    1   0   1   0   1   1

; LR: a`gree+ed a`gree+ing a`gree+able `hoe+ed `hoe+ing `dye+ed `dye+ing
```

```
RULE
"y:i => @:CN ___ +:@ ~[i | ']"  4 6
     @   y   +   i   '   @
     CN  i   @   i   '   @
 1:  2   0   1   1   1   1
 2:  2   3   2   1   1   1
 3.  0   0   4   0   0   0
 4.  2   1   1   0   0   1

; LR: `trace+able  `change+able
; SR: OtraceOable  OchangeOable
;Elision prohibited after CNpal and before
back vowel
RULE
"e:0 /<= VO CN* CNpal ___ +:@ VObk" 5 7
      VO  CN  CNpal  e  +  VObk  @
      VO  CN  CNpal  0  @  VObk  @
 1:   2   1     1    1  1    2   1
 2:   2   2     3    1  1    2   1
 3:   2   1     1    4  1    2   1
 4:   2   1     1    1  5    2   1
 5:   2   1     1    1  1    0   1
```

```
; SR: aOgreOOed aOgreeOing aOgreeOable OhoOOed OhoeOing OdyOOed OdyeOing
RULE
"e:0 <= CN [VO | y] ___ `+:@ e" 5 8
     CN VO y  e  e  +  `  @
     CN VO y  0  @  @  0  @
 1:  2  1  1  1  1  1  1  1
 2:  2  3  3  1  3  1  2  1
 3:  2  1  1  1  4  1  3  1
 4:  2  1  1  1  1  5  4  1
 5:  2  1  1  1  0  1  5  1
```

```
; LR: `argue+ing `argue+able
; SR: OarguOOing OarguOOable
RULE
"e:0 <= CN u ___ +:@ VO" 5 7
     CN u  e  e  +  VO @
     CN u  0  @  @  VO @
 1:  2  1  1  1  1  1  1
 2:  2  3  1  1  1  1  1
 3:  2  1  1  4  1  1  1
 4:  2  1  1  1  5  1  1
 5:  2  0  1  0  1  0  1
```

```
;clean-up rule for three <= Elision rules plus i:y rule
RULE
"e:0 => [ VO CN CN* ___ +:@ VO | CN [VO | y] ___ +:@ e | CN u ___ +:@ VO | i:y ___ +:@ i]" 14 11
      e   @   CN  +   e   y   u   i   `   i   @
      0   VO  CN  @   e   y   u   i   0   y   @
 1:   0   2   6   1   2   6   2   2   1   12  1
 2:   0   2   3   1   2   3   2   2   2   12  1
 3:   4   7   3   1   7   3   11  7   3   12  1
 4.   0   0   0   5   0   0   0   0   4   0   0
 5.   0   2   0   0   2   0   2   2   5   0   0   ;prohibits e:0 before y suffix
 6:   0   7   6   1   7   10  11  7   6   12  1
 7:   8   2   3   1   2   3   2   2   7   12  1
 8.   0   0   0   9   0   0   0   0   8   0   0
 9.   0   0   0   0   2   0   0   0   9   0   0
10:   8   7   6   1   7   10  11  7   10  12  1
11:   4   2   3   1   2   3   2   2   11  12  1
12:   13  1   6   1   1   6   1   1   12  12  1
13.   0   0   0   14  0   0   0   0   0   0   0
14.   0   0   0   0   0   0   0   1   0   0   0
```

```
;==========
;i:y-spelling
;==========
;die+ing  tie+ing
;dyOOing  tyOOing
;e:@ instead of e:0 to force Elision to apply
RULE
"i:y <= ___ e:@ +:0 i" 4 6
     i  i  e  +  i  @
     y  @  @  0  i  @
 1:  1  2  1  1  2  1
 2:  1  2  3  1  2  1
 3:  1  2  1  4  2  1
 4:  1  0  1  1  0  1
```

```
RULE
"i:y => ___ e:0 +:0 i" 4 5
     i  e  +  i  @
     y  0  0  i  @
 1:  2  1  1  1  1
 2.  0  3  0  0  0
 3.  0  0  4  0  0
 4.  0  0  0  1  0
```

```
;==========
;Gemination
;==========
; LR: re`fer+ed  `travel+ed  `travel+ed  `sleep+ing  `slip+ing  `spot+y  `spot+y+ness
; SR: reOferred  OtravelOed  Otravelled  OsleepOing  Oslipping  Ospotty  OspottiOness
RULE
"+:{b,d,f,g,l,m,n,p,r,s,t} /<= `:0 CN* VO {b,d,f,g,l,m,n,p,r,s,t} ___ [VO | y:@]" 5 17
      `   CN  VO  b   +   y   d   f   g   l   m   n   p   r   s   t   @
      0   CN  VO  b   0   @   d   f   g   l   m   n   p   r   s   t   @
 1:   2   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 2:   2   2   3   2   1   1   2   2   2   2   2   2   2   2   2   2   1
 3:   2   1   1   4   1   1   4   4   4   4   4   4   4   4   4   4   1
 4:   2   1   1   1   5   1   1   1   1   1   1   1   1   1   1   1   1
 5:   2   1   0   1   1   0   1   1   1   1   1   1   1   1   1   1   1
```

```
;correctly permits traveled and travelled,
; but also incorrectly permits gemination in affixes (see next two rules)
RULE
"+:{b,d,f,g,l,m,n,p,r,s,t} => CN* VO {b,d,f,g,l,m,n,p,r,s,t} ___ [VO | y:@]" 15 27
      +  +  +  +  +  +  +  +  +  +  +  CN  VO  b  +  y  d  f  g  l  m  n  p  r  s  t  @
      b  d  f  g  l  m  n  p  r  s  t  CN  VO  b  @  @  d  f  g  l  m  n  p  r  s  t  @
 1:   0  0  0  0  0  0  0  0  0  0  0  2   1   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 2:   0  0  0  0  0  0  0  0  0  0  0  2   3   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 3:   0  0  0  0  0  0  0  0  0  0  0  2   1   4  1  1  6  7  8  9  10 11 12 13 14 15 1
 4:   5  0  0  0  0  0  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 5.   0  0  0  0  0  0  0  0  0  0  0  0   1   0  0  1  0  0  0  0  0  0  0  0  0  0  0
 6:   0  5  0  0  0  0  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 7:   0  0  5  0  0  0  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 8:   0  0  0  5  0  0  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
 9:   0  0  0  0  5  0  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
10:   0  0  0  0  0  5  0  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
11:   0  0  0  0  0  0  5  0  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
12:   0  0  0  0  0  0  0  5  0  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
13:   0  0  0  0  0  0  0  0  5  0  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
14:   0  0  0  0  0  0  0  0  0  5  0  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
15:   0  0  0  0  0  0  0  0  0  0  5  2   2   2  1  2  2  2  2  2  2  2  2  2  2  2  1
```

```
RULE
"no gemination in prefixes" 2 6
     #  -  +  +     `  @
     #  -  0  CNgem  0  @
 1:  2  2  1  1     1  1
 2:  1  2  2  0     1  2
```

```
RULE
"no gemination in suffixes" 3 4
       `  +  +      @
     0  0  CNgem  @
 1:  2  1  1      1
 2:  2  3  3      2
 3:  3  3  0      3

;==========
;s-deletion
;==========
;These rules handle deletion of the possessive s:
; LR: cat+s+'s  fox+s+'s  Dallas+'s
; SR: cat0s0'0  foxes0'0  Dallas0'0
;If you don't want to use these rules, move them after the END keyword and
; in the file affix.lex, comment out the +'s entry and uncomment the +' entry.
;These rules increase recognition time.

RULE                                                    RULE
"s:0 => s +:@ ' ___" 4 5                                "s:0 <= +:@ s +:@ ' ___" 5 5
      s  s  +  '  @                                         s  s  +  '  @
      0  s  @  '  @                                         0  @  @  '  @
 1:   0  2  1  1  1                                    1:  1  1  2  1  1
 2:   0  2  3  1  1                                    2:  1  3  1  1  1
 3:   0  2  1  4  1                                    3:  1  1  4  1  1
 4:   1  2  1  1  1                                    4:  1  1  1  5  1
                                                       5:  1  0  1  1  1
END
```

Az angolnál nincs szükség kiegészítő karakterekre. A teljes morfológia leírás nem is nagy. A magyarra nem a legalkalmasabb a módszer. A szótárban kiegészítő kódokat kell alkalmazni, vagy a konkatenációs lehetőségekkel kell megoldani olyan osztályozásokat, ami pusztán a szó alakjából nem egyértelműen meghatározott. A finn nyelv morfotaktikája hasonlóan összetett, mint a magyaré, de a fonológiája következetesebb, mert a szavak, morfémák alakjai meghatározzák a fonológiai szabályokat. Itt sincs szükség kiegészítő karakterekre a szótári szinten.

A módszer átütő volt. Számos nyelvre születtek implementációk. Köztük az inflexiós tulajdonságokkal rendelkező arabra is. Egyik pozitívum, hogy a különböző nyelvi jelenségeket függetlenül kell leírni. Az ennek megfelelő kis automaták párhuzamos futtatása nem sok helyet követel. A formalizmus egyszerű. Ha hasonló fonetikai eszközzel találkozott egy nyelvész, könnyedén írhatja le az egyes jelenségeket. További előnye (és esetleg hátránya), hogy egy leírás alkalmas elemzésre, generálásra egyaránt. Ilyen értelemben helyesírás-ellenőrzőnek is jó, és alkalmazzák is. A leírás alapján képesek determinisztikus módon viselkedni, abban az értelemben, ha a két input, a felszíni és a szótári input megfelelőségét kell ellenőrizni. Azonban sem generátorként, sem elemzőként nem determinisztikus. Ezt nem is várhatjuk, hiszen ebben az értelemben nem feltétlen egyértelmű a működése.

A leírási módnak van néhány korlátja. Az egyik, hogy a formalizmusban nincs helye annak, hogy két karakter azonos. Ez a magyarban, az angolban is a betűkettőzés leírásánál mutatja hátrányát. Mindkét esetben kihagyták a teljes hasonulásnak – a gemination-nek formális leírását, helyette praktikusabb egyből az automatát megszerkeszteni. A másik korlát, hogy a távolsági összefüggések, kezelésére nincs praktikus módszer. A folytatási osztály használata sokszorozná a lexikon méretét. A magyarban az igekötők és felsőfok engedélyezését segédkarakterekkel oldottuk meg. Vannak módszerek, melyek az ilyen leírásból akár determinisztikus helyesírás-ellenőrzőt készítsen automatikusan. A magyarra ez teljes kudarcba fulladt a nagy állapotszám miatt.

Érdekes, hogy az elnyelődést több implementáció csak a szótár-felszín irányba engedi meg. Formálisan nincs különbség a két szint között, ennek ellenére az `ε:x` nem megengedett, csak az `x:`ε. Persze kérdés, lehet funkció nélkül egy felszíni karakter.

4.12.6 Egy gyors unifikációs módszer (HUMOR)

A korábbi algoritmusok mind balról jobbra működtek, pontosabban elölről a vége felé, ne feledkezzünk meg azokról, akik más irányban írnak. A reguláris nyelvek tükrei is regulárisak, így megpróbálhatjuk a szavakat a végéről elemezni. Magyarban erre időnként szükség is lehet. Ha ismeretlen szó kerül a horogra, akkor is szükség lehet elemzésre. Ha azt írom, hogy *Barack Obamával*, akkor szótár nélkül is rá lehet jönni, hogy egy név szerepel a szövegben *-val* raggal. Ezt viszont nem érdemes balról jobbra elemezni, mert így minden pozícióban kereshetem a toldalékot. Talán érdemes lenne egyszer megpróbálni a magyar szavakra olyan végesállapotú automatát gyártani, ami jobbról elemez. A toldalékokat viszont jobbról toldalékoló nyelveknél mindenképp érdemes jobbról vizsgálni.

Ezt használja ki a HUMOR eredeti implementációja. Ha elfeledkezünk a prefixekről, szóösszetételekről, akkor a szó két részből áll: egy szótőből és egy toldalékszekvenciából. A szótövet a szó elejétől keressük, a lehetséges toldalékokat viszont a szó végéről. Ha egy kezdet lehet szótő, egy vég lehet végződésszekvencia, akkor, ha a kettő egy pontban találkozik, szerencsénk van, de ellenőrizni kell, jól csatlakozik-e a kettő. Így is csinálta az 1993-as implementáció. Nagy előny, hogy csökkenti a felesleges próbálkozásokat, az elemzés holtágait. Ráadásul a tő és a toldalék találkozásánál egy egyszerű tulajdonságösszevetés, unifikáció dönti el, vajon igazi csatlakozás, vagy véletlen téves próbálkozásról van szó:

b betű	b							
a vitorlás rúdja	b	o	c					
elnézés vagy medvefia	b	o	c	s				
elnézés	b	o	c	s	á	n	a	t
pozíció	1	2	3	4	5	6	7	8
nincs toldalék								
tárgyrag vagy múlt idő								t
tárgyrag vagy műveltetés							a	t

Az adott szónál, ha balról jobbra mennénk, akkor 4 helyen próbálkoznánk a toldalékokkal. Ebből a *bocs* után addig, hogy *ána* még kecsegtethet sikerrel a toldalékszekvenciák között, csak utána derülne ki, hogy felesleges volt a próbálkozás. Ehelyett a HUMOR csak a végén, a 8. pozícióban talál szótő-toldalék találkozását. Az ellenőrzés (üres toldalék) itt abból áll, hogy a megtalált szótőforma megáll-e önmagában, vagy sem.

k betű	k					
főnév	k	e	r	t		
pozíció	1	2	3	4	5	6
nincs toldalék						
birtokrag						m
birtokrag vagy igerag					e	m
igerag				t	e	m

Ebben az esetben csak a 4. pozícióban találkozik jobbról egy toldalék balról egy szótővel.

A két rész egyesítése (unifikálása) akkor lehetséges, ha a kiválasztott szótőforma és toldalékforma tulajdonságai megengedik. A HUMOR-ban ez két részből áll. Ezekről a **4.10.13**-as fejezetben volt szó. A vizsgálat egyik tárgya, hogy a szótő betűjele és a toldalék betűjele konform-e – egy nem nagy mátrixban tárolja a rendszer, hogy milyen tőformatípusok milyen toldalékfajtákkal illenek össze. A másik vizsgálat főleg egyszerű fonetikai tulajdonságokat ellenőriz, melyek binárisan vannak tárolva. Az ellenőrzés egy egyszerű maszkolt logikai művelet a két bitmintán. Szinte semmi időbe nem telik.

4.13 Gyakorlati kérdések

4.13.1 Helyesírás-ellenőrzők

A helyesírás-ellenőrzőnek két szerepe van.

1. Felismerje a hibásan írt szavakat
2. Segítse a szövegjavítást

Az ellenőrzők algoritmusa többnyire környezet nélkül, csupán önálló szavakat vizsgál. Vannak ugyan kísérletek arra, hogy ennél többre legyen képes, de ezek eredményei kétesek, illetve túlmutatnak a szóellenőrző lehetőségein. Ilyen, ha egy szónál érzékeljük, mondat elején szerepel vagy sem, és eszerint engedélyezzük, hogy nagy kezdőbetűvel írják. A mondat elejének megállapítása viszont nem egyszerű feladat, és ha túl egyszerű az algoritmus, gyakran tévedhet a szövegszerkesztő ellenőrzője. A mondat elejének megtalálásában az egyik fő ismérv, hogy nagybetűvel kezdődik. Másrészt néha olyan szövegeket kell ellenőrizni, melyben nem teljes mondatok vannak, tehát az első szót sem kell nagybetűsíteni.

Gyakori, ha a szövegben szóismétlés esetén hibát jeleznek az ellenőrzők. Nos, angolban is előfordul ilyen, magyarban pedig nem ritka az *az az* rész, ahol az első *az* vonatkozó névmás, a második pedig névelő – tehát nem igazi szóismétlés. Más szónál is előfordulhat látszólagos szóismétlés (*vagy*, *is*. *mert*), de nem gyakori.

Ez ezt követi, az az az után sétáló embert.

Maradjunk a szó izolált ellenőrzésénél. Ha igazi szöveget vizsgálunk, a kis-nagybetű váltást az első karakternél engedélyezni kell. Hasonlóan a csupa nagybetűvel való írást is. Ennek algoritmizálása (nem generátor jellegű, hanem a felszíni alakról kell mondani valamit) növeli az algoritmusok indeterminisztikus voltát. Angol nyelvnél még az a luxus is megtehető, hogy on-line legeneráljuk a szavak mindhárom változatát, és az így kialakult háromszoros méretű szótárból készítjük el a szpeller automatáját. A legtöbb nyelvnél ez nem járható út a korábban említett méretek miatt.

A következő gond ott van, hogy ha a szótant le is tudnánk írni teljes pontossággal, a szótár sohasem lehet teljes. A szótárnak mindig nyíltnak kell lennie. Ha más nem, nevek lehetnek a szövegben, de az élő nyelv mindig produkál új szavakat. Ha azt nézzük, hogy a jelenlegi szpellerek hány százalékát találják meg a szöveg szavainak, akkor a jó minőségűek is csak 95-99 %-át ismerik fel. Ráadásul e szám fölé nem is lehet eljutni. Ismervén a Zipf törvényt és az azt, hogy az adatbevitel is lehet hibás, a ritka szavaknak csak kis hányadát tudjuk lefedni. Egy nyelvi eszköz a nyelvet X %-ban fedi le, ha a futó szöveg X %-át ismeri fel jól.

lefedettség=(helyesnek értékelt szavak száma a helyesek közt)/(helyes szavak száma)

Ha nem ismerjük a helyes szavak számát, csak a helyesnek értékeltekét, akkor a következő arányt lehet megadni, amit **recall**-nak hívnak. Erre nem ismerek jó magyar szót, talán a **visszaigazolás** lenne megfelelő.

visszaigazolás=(helyesnek értékelt szavak száma a helyesek közt)/(helyesnek értékelt szavak száma)

Hiú remény, hogy egy helyesírás-ellenőrző 99%-os lefedettséget produkáljon. Pontosabban, ha például mindent elfogad, akkor 100 %-os a lefedettség, de a nagyobb hiba, hogy hány rossz szóra adja azt a választ, hogy helyes, az siralmas lenne. E másik paraméter, hogy hány helytelen szót fogad el. Ezt én **mellényúlás**nak nevezném.

mellényúlás=(helyesnek értékelt szavak száma a helytelenek közt)/(helytelen szavak száma)

Nem ismerek jobb magyar szót rá. Nos, a mellényúlás egy fontosabb dolog, mint a lefedettség. Az ellenőrző pontossága e kettőből tevődik össze.

pontosság=(helyes döntések száma)/(összes előforduló szó száma)

ahol

(helyes döntések száma)=(elfogadottak száma a helyesek közt)+(elutasítottak száma a helytelenek közt)

(összes előforduló szó száma)=(helyesek száma)+(helytelenek száma)

A felhasználó szempontjából az a nagyobb hiba, ha hibás szót fogad el az ellenőrző. Az ilyen szavak kicsúsznak az ellenőrzés alól, míg a másik hiba esetén, ha helyes szóra csenget a rendszer, akkor a felhasználó még mindig felülbírálhatja az ellenőrzőt. Tehát ha súlyozni kell, milyen jó egy helyesírás-ellenőrző, akkor

minőség=α(lefedettség)+(1-α) (mellényúlás)

ahol α egy kis pozitív szám. Hogy mekkora, az már ízlés dolga. Attól függ, ki mennyire taksálja az egyik, illetve a másik hibát, de 0,1-nél nem érdemes nagyobbra venni.

Gyakorlat, hogy munka közben a felhasználó bővíthesse szótárát. Ha talál egy hibásnak jelzett szót az ellenőrző, akkor a gép döntését felülbírálva akár rögzítheti: ha még lesz ilyen, akkor azzal már ne kelljen törődni. Az így készített felhasználói kiegészítő szótár dinamikusan változik. Méretét ugyan nem szokták korlátozni, de nem akkora, mint az általános nagyszótár. 10 000 tételnél csak a profik használnak nagyobbat, azt is csak az olyan nyelv esetén, mint a magyar. Emiatt tárolása nem kritikus, de ha gazdaságosan és gyorsan kereshetővé akarjuk tenni, akkor lehet szófában reprezentálni.

Megjegyzem, hogy egy felhasználó elérhet 99 %-nál nagyobb pontosságot. A magyarázat az, hogy ekkor nem a magyar nyelvet kell lefedni, hanem a sajátját, ami egy sokkal kisebb halmaz. Ezért a ritka szavak nagy hányada is bekerülhet a saját szótárába. Nos, ezeknek a felhasználóknak sem éri el a 10 000-et a méret, csak akkor, ha a felhasználó sok forrásból, több szerzőtől kapott anyagokat ellenőriz. Magyarban viszont még ez sem segít, mert egy szónak is sokezer előfordulási alakja van. Emiatt ahhoz, hogy a felhasználói szótár lényegesen javítson a lefedettségen, már akkora méret szükséges, aminek mérete az egy ember által karban tarthatót túllépi. Ezen úgy lehet segíteni, ha egy felvett szóval automatikusan a szó más előforduló formáit is kezelné a szóellenőrző. Erre vannak kísérletek a HUNSPELL-ben és a HUMOR-ra épülő Helyes-e? programban. Mivel egy átlagos felhasználó még az egyszerűbb kérdésekre sem tud jól válaszolni a szó fonológiai osztályozásánál, a megoldás az, hogy meg kell adni a rendszer által ismert készletből egy mintaszót, és jelezni: a *kozmál* szót toldalékold úgy, mint a *hátrál* szót. Ez a megoldás sem vált be igazán, a tömeges felhasználók nem éltek vele. A profi felhasználóknál viszont sokat segít.

Ha egy szó a nyelvi modell szerint helyes, akkor nem biztos, hogy jó, ha elfogadja az ellenőrző. Mint korábban kifejtettem, a nyelvi rendszerbe is építenek olyan eszközöket, melyek bizonyos elemzéseket nem fogadnak el akkor sem, ha helyes, hogy csökkentsék a többértelműséget. Az ellenőrzőnél akkor is ajánlatos egy elemzést elvetni, ha a szónak ugyan lehet a modell szerint elemzése, de egyrészt sohasem használják, másrészt nagyon közel van gyakran előforduló szóhoz, amihez „közel” van, tehát gyakori írási hibával bekerülhet a szövegbe. A *tanit*=*tan*[FN]+*i*[IKEP]+*t*[ACC] helyes szó, de biztosan a *tanít* ige elírásából ered.

Ha lehetőség van új szavak lajstromba vételére, akkor természetes, hogy szavak tiltását is meg kell engedni. Soha nem lehet tudni, miről feledkeztek meg a nyelvi információt előállítók. Míg a felhasználói szótár a lefedettséget növeli, a tiltószótár a mellényúlást csökkenti. Persze ezek a kiegészítések formálisan párhuzamos egységek, ezért jól meg kell határozni, mi az elfogadandó szó. Ha az általános ellenőrző nyelve az L_g,

a felhasználói szótáré az L_u és a kivételszótáré az L_e, akkor az elfogadott nyelvé $L=(L_g+L_u)-L_e$ amiből következik, hogyha egy szó le van tiltva, akkor azt akkor is elutasítja az ellenőrző, ha felveszi a saját felhasználói szótárába.

Feladat 1: Végezzen kereszttesztet azonos nyelv különböző szpellerével.

Feladat 2: A tesztet úgy is végezze el, hogy generált szövegen teszteljen.

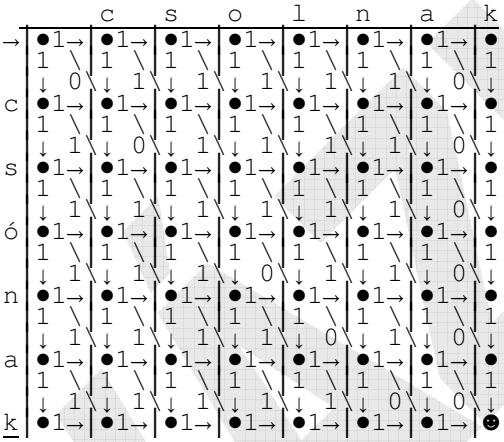
4.13.2 Korrekciós algoritmusok

Az ellenőrzők másik fontos feladata az ajánlás. Ha hibás szót talál a program, akkor ajánl egy vagy több másikat helyette. Létezik automatikus javító és interaktív módszer. Mindkét esetben lényeges, hogy a gép kitalálja, mi lehetett a szó eredeti formája a hiba elkövetése előtt, ha helytelen alakot talál. Az alapalgoritmus feltételezi, hogy az elírások elemi független hibákból tevődnek össze, és a hibák (súlyainak) összege adja az összhiba mértékét.

A legegyszerűbb hibasúlyozás a **Lemming-távolság**. Két sztring szomszédos (távolságuk 1) ha egyiket a másikból úgy kapjuk, hogy

1. egy karaktert kicserélünk egy másikra, vagy
2. egy karaktert kihagyunk, vagy
3. egy karaktert beszúrunk

Az ebből a szomszédsági relációból indukált legrövidebb út hossza a Lemming-távolság. Két szó távolságához, ha egyik `n`, a másik `m` hosszú, egy `(n+1)×(m+1)` méretű gráfban keresendő a minimálút. Az ebből a négyzetből „kilógó” utak nem lehetnek minimálisak. Sőt, az éleket lehet irányítottnak venni.



Forrás: `csolnak` Cél: `csónak`

→ törlés – `1` pont

↓ beszúrás – `1` pont

\ Ha azonos a két karakter, továbblépés – `0` pont, egyébként helyettesítés – `1` pont

A minimálköltségű útra sok jó algoritmus létezik. Az algoritmusok viszont a gráfpontok négyzetével arányos lépésszámban oldják meg a feladatot. Kihasználva a gráf ciklusmentességét, az élek nem negatív voltát és az egyszerű mátrixos elrendezést, **Vlagyimir Levenstein** a csúcsokkal arányos algoritmust ad meg. Egyszerűen sorról sorra végigmegy az `(n+1)×(m+1)` méretű mátrix elemein, és azt a legfeljebb három közvetlen szomszédból induló élek mentén (az eggyel kisebb indexű) elemeket vizsgálja, melyekre már korábban kiszámította, milyen költséggel lehet eljutni a bal felső sarokból. A lépésszám nyílván lineáris a csúcsok számával.

		c	s	o	l	n	a	k
	0	1	2	3	4	5	6	7
c	1	0	1	2	3	4	5	6
s	2	1	0	1	2	3	4	5
ó	3	2	1	1	2	3	4	5
n	4	3	2	2	3	2	3	4
a	5	4	3	3	4	3	2	3
k	6	5	4	4	5	4	3	**2**

Ha az érdekel, hogy egyik szóból korlátozott módosítással el lehet-e jutni a másikba, akkor a sebesség növelhető. Ha például `2`-nél több korrekció esetén már nem érdekel a távolság, mert az ennél nagyobb eltérés esetén távolinak tartjuk a két szót, akkor a mátrix főátlójától három éllel messzebb eső pontokat nem is kell vizsgálni. A főátló nem matematikai értelemben értendő, hiszen a mátrix nem feltétlen négyzetes. Itt azt jelenti, hogy ha i és j a mátrix elemeinek indexe, akkor $|i-j|\leq 2$ feltételnek eleget tevő elemekkel kell csak számolni. Ellenkező esetben az átmenetben vagy beszúrásból, vagy kihagyásból legalább `2`-re lenne szükség. Ezzel az algoritmus hely és időigényét nem a mátrix méretével, hanem a két sztring hosszával lehet arányossá tenni.

		c	s	o	l	n	a	k
	0	1	2					
c	1	0	1	2				
s	2	1	0	1	2			
ó		2	1	1	2	3	4	
n				2	3	2	3	4
a					4	3	2	3
k						4	3	**2**

Az algoritmus azzal is pontosítható, ha a különböző kihagyásoknak, beszúrásoknak, helyettesítéseknek egytől különböző súlya van. Például, ha szövegszerkesztő használatáról van szó, akkor jellegzetes elütések (egymás melletti billentyűk, rövid *i* hosszú *í*) kis súllyal, a ritkák nagy súllyal számíthatók. Egy karakterfelismerésnél a hasonló formájú karakterek távolsága kisebb, mint az egymásra egyáltalán nem hasonlítóké.

Az alapvető gond nem is ott van, hogy az elemi távolságot hogyan számítsuk. A Levenstein-algoritmus általánosítható ezekre is. A baj ott van, hogy ha forrásnak tekintjük a leírt szót, a cél nem egy kiválasztott másik szó, hanem a magyar nyelv azon szavának kiválasztása, amelyik legjobban hasonlít rá – a távolsága legkisebb a forrástól – vagy legalábbis egy halmaz azon szavait keressük, melyek közel vannak a forráshoz, akkor minden halmazbeli szót össze kéne vetni a forrással.

Emiatt inkább az az eljárás, hogy a forrásszón megkísérlünk egyszerűbb korrekciókat végezni, majd az így kapott jelöltet vetjük alá a helyesírás-ellenőrzőnek. Ebben az esetben, ha csak egyfajta elemi javítást engedünk meg korrigálni, és n hosszú a szó, N az abc mérete, akkor $n\times(2N+2)$ próbát kell megejtenünk, vagyis több ezerszer kell lefuttatni az ellenőrzőt. Ha megengedjük a kettős hibát, akkor sokmilliónyi a próbálkozások száma. Ez még a mai gyors ellenőrzőknél is túl sok időbe telik.

A legegyszerűbb, ha csak a legvalószínűbb korrekciókat próbáljuk. Vagyis nem 1 a súlya minden betűcserének, hanem az alkalmazásnak megfelelően csak a nagy valószínűségű hibák korrekcióját engedjük kijavítani. Ebből viszont kettőt, hármat is. Ha egy pozícióban a javítási lehetőségek száma korlátozott, például 5, akkor egy átlagos szónál a lehetséges korrekciók száma 1000 alatt marad kettős hiba esetén is.

Ha egy elemi i hiba elkövetésének valószínűsége P_i, és a hibák elkövetése független, akkor annak a valószínűsége, hogy S sztringből $\not{S}$-t kapjuk i_1, i_2… i_k hiba következtében, $P(S\to\not{S})=\prod P_i$ Sajnos nem ismerjük a jó szót, hanem a hibást. Emiatt azt az S szót keressük, amire a $P(S\to\not{S})$ maximális, vagy azokat melyeknél a fenti valószínűség egy határérték felett van. Erre megy ki a korábban említett algoritmus.

Az elvi megoldás nem ez. Ha az a kérdés, milyen szó helyett írták le a hibás, esetleg hibátlan sztringet, akkor a következő értékeket kell összehasonlítani $P(S)\times P(S\to\not{S})$ tehát szorozni kell az elírás $P(S\to\not{S})$ valószínűségét annak az S szónak a $P(S)$ valószínűségével, amelyet (esetleg) elírtak. Legalábbis, ha a szó előfordulása független, vagy legalábbis korrelálatlan az előforduló elírási típussal. Ebből lehet naiv Bayes-módszer alapján számolni, milyen szó a valószínűsített szándék.

Az elmélet szép, a gyakorlatban a szavak többségének valószínűségére megbízhatatlan becslések vannak a szövegenkénti nem egységes eloszlások miatt, de szerzőnként is lényeges eltérések lehetnek. Stáhl Judit szakácskönyvének első felében például a *só* szó gyakoribb, mint a másodikban, és a *cukornál* fordítva, a másodikban a gyakoribb, mint az első részben. Az ok egyszerű, az édességek a kötet második felébe kerültek. Más szövegkörnyezetben más a valószínűség. Ennek ellenére már alkalmazzák a szógyakoriságot az ajánlásnál. Amit fontos lenne figyelembe venni, hogy két kis tapasztalati gyakoriságú szónál már nem lenne szabad, hogy a globális szógyakoriság befolyásolja a döntést, mert az ebből számított valószínűség megbízhatósága alacsony. Ennek ellenére hasznos ez a megközelítés. Ilyen ajánlója van a FireFoxnak, a Google-nak, és sok más programnak. Néha teljesen mellélő – olyan gyakori szavakat ajánl korrekcióként, melynek nagyon kevés alaki köze van a beírthoz.

A közeli szóformák keresésének bevált általános módszere az n-grammok statisztikájának felhasználása. Itt nem a teljes szavak gyakoriságát tárolják – ennek megbízhatósága nem elég – hanem általában 3-grammokét. Erre viszont jobb statisztikát lehet készíteni, mert számuk sokkal alacsonyabb, mint a szóalakoké. Ennek alapján próbálják a valószínű elütéseket megtalálni. Minden pontban a szomszédos 2 karakter alapján a statisztika sugallatára pár módosítással próbálkoznak, így – hasonlóan a jellegzetes elírásokra támaszkodó módszerhez – szavanként párezer ellenőrzéssel keresik a lehetséges elírást.

Ennél sokkal jobb módszer, ha a helyesírás-ellenőrző motorjába belenyúlunk, és kihasználjuk azt, hogy ellenőrzés közben sokkal több információt nyerhetünk ki, mint csak azt, hogy helyes-e egy szó. Ha az ellenőrző balról jobbra halad az elemzendő sztringen, mert például véges automata a szótan reprezentációja, akkor azt is meg lehet tudni, melyik volt az a pozíció, aminél végképp leakadt az elemzés. Ha korrekciónál csak e hely után javítanánk a szövegen, biztos hogy elfogadhatatlan szót kapunk. Ennek ismeretében általában felezhetjük az ellenőrizendő javítások számát. Ennél sokkal többet segít, ha minden pozícióban letároljuk, milyen karakterekre léphetünk tovább (nem csapda állapotba) az automatánkkal. Természetesen indeterminisztikus automatánál ez kumulatív információ minden bejárt elemzési kísérletnél. Nos, ha ezt is felhasználjuk, akkor nagyságrendekkel lecsökken az ajánlásoknál az igazából ellenőrizendő szójelöltek száma akármilyen javítóalgoritmust választunk a korábban jelöltekből. A gyakorlatban ez a módszer a legha-

tékonyabb, de persze ehhez szükséges, hogy az elemzés véges automata elvén (esetleg szófával) történjék, és az elemzés alatt képes legyen kigyűjteni a program a bejárt csomópontokból kiinduló nem használt éleket is.

Feladat 1: Tegyük fel, hogy a hibák súlyozása 1:10, vagyis a mellényúlás tízszer akkora, mint a lefedettségből származó hiba. Hogyan állapítaná meg, hogy az álnak szót elfogadja-e a helyesírás-ellenőrző?

Feladat 2: Ha lehetne szókörnyezetet is vizsgálni, mondjuk közvetlen szomszédságot, akkor hogyan csökkenthetné a mellényúlást úgy, hogy a lefedettség nem csökkenne.

Feladat 3: Keressen olyan szavakat (szóformákat) különböző nyelveken, amelyek egymást követően ismétlődhetnek szabályos szövegben.

4.13.3 Automatikus szöveg-helyreállítások

A hibák javítása automatikusan is történhet. Ekkor a korábban említett valószínűségi meggondolások alapján tehetjük meg. A kísérletek angol nyelvre hasznosnak is mutatkoznak. Ha jó súlyozást alakítunk ki a jelöltek között, akkor szövegszerkesztőknél legalább 70 százalékban biztonsággal meg lehet saccolni a korrekciót. Emiatt a hibafelderítések negyedénél 95 százaléknál jobb automatikus javítást lehet elérni. A tesztek sem magyar, sem román nyelvnél nem mutattak bíztató eredményeket. A románnál még elképzelhető, hogy a súlyozási algoritmus nem volt megfelelő, de magyarban annyira sűrű a szavak tere, hogy reménytelen 50 százalékosnál jobb korrekciós eredményt elérni.

Erre utal az **ékezetesítő** programnál tapasztalt eredmény is. Mivel gyakran fordul elő, hogy üzeneteket kapunk éketlen írással, jelentősége van ennek is. A program csak arra vállalkozik, hogy a *vereb* szóra a *véreb* és a *veréb* szót is megleli, de választani a felhasználónak kell. Persze az *és* kötőszó az *es*-ből egyértelműen rekonstruálható.

Ehhez hasonló program a **zárt *ë*-k rekonstruálásá**ra készített szolgáltatás is. A programnak az a feladata, hogy helyesen írt magyar szövegben különböztesse meg a kétféle *e*-t. Ez az esetek többségében egyértelmű, de időnként – az alternatív szóelemzések esetén csak humán beavatkozás dönthet. A statisztika persze itt is segíthet, de félre is vezethet.

A sebesség a hétköznapi felhasználásnál ma már nem izgalmas. Egy szövegszerkesztőnél, ha a korrekcióra kérdezünk interaktív üzemmódban, akkor az egy másodpercen belüli válaszidő elfogadható. A fenti gyorsításos módszereknél, még ha párezerszer kell végrehajtani próbát, akkor is az időn belül maradunk. Más felhasználásnál viszont nem így van. Beszéd-, illetve optikai szövegfelismerésnél a hang illetve a képi információból sokkal gyakoribb a bizonytalanul felismert betű – sőt sokszor a szóhatár is kérdéses. A válaszidő viszont korlátozottabb, hiszen például beszédfelismerésnél a beszéddel szinkronban illik megadni a választ. A klasszikus felismerők megadják, miben biztosak, de amiben bizonytalanok, ott csak azt az információt adják át, hogy ott is van egy-két betű. Pedig a felismerők ennél többet hallanak, látnak. Például, hogy egy balra gömbölyű hasú karakter, vagy egy vagy két zöngés, fognál képzett mássalhangzó. Nos, ha ezt az információt átadná a korrektornak, akkor nagyobb eséllyel található ki, mi lehet az szó. Ha ráadásul valószínűségeket is rendel az optikai vagy hangfelismerés felismerés alternatíváihoz, nyert ügyünk van. Ezek az alternatívák kizárólagosan lokális darabkára vonatkoznak, tehát véges lehetőségek közt lehet a megoldás. Ezt valószínűséggel kiegészített véges (ciklusmentes) automatával reprezentálni lehet. És ezek után a két automatát, a nyelvit is a felismerő által reprezentált inputot kell paralel futtatni, az összvalószínűséget kiértékelni.

Azt fontos tudni, hogy míg a korábbi alkalmazásoknál a futó szöveg szavainak többsége helyes, az optikai és hangfelismerésnél az ellenőrizendő alternatívák többsége nem fogadható el magyar szónak. Emiatt is kell más logikát követnie ezeknek az algoritmusoknak. Ha egy szóban a karakterfelismerő 3 helyen bizonytalan, ha csak a bizonytalanság információt kapná meg az automatikus korrektor, akkor körülbelül 10 000 szóalakot kéne kipróbálni. Ha viszont a bejövő információból lehet következtetni a bizonytalan karakterre, helyenként 10 alatti lehetőség maradna, akkor 1000 szóalakot kell csak ellenőrizni. Ha ráadásul a hang-, karakterfelismerő eredményét megfelelő automatával (reguláris kifejezéssel) prezentáljuk, akkor pillanatok alatt számítható a lehetséges megoldás még sok bizonytalanság esetén is. Ha pl. a következőt ismeri fel az OCR:

```
<high><:><high><:><under><under><mrn>i<low>i<low2><high><'><low>i<low2.5>
```

ahol
- `<:>` umlautos magánhangzó [öőüű]
- `<'>` egyékezetes hosszú magánhangzó [áéíló|ú]
- `<high>` magas karakter [b|d|f|h|k|l|t|A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z…]
- `<under>` lelógó karakter [g|y|p|q]
- `<mrn>` m vagy rn (tipikus OCR probléma) [m|(rn)]
- `<low>` alacsony normál széles karakter
- `<low2>` egy széles vagy két normál alacsony karakter

`<low2.5>` két vagy három alacsony karakter

akkor a keresett szó: `Külügyminisztérium` lehet csak, és annak ellenére, hogy 13 bizonytalanság van a karakterfelismerésnél, ez századmásodperc alatt tisztázható.

A beszédfelismerésnél a szavak különválasztása okoz többletgondot: attól, hogy szünet nélkül mondunk egy kifejezést, még sok szóból is állhat a beszédrészlet. Itt a többértelműség lényeges forrása a szóköz felismerése. Ráadásul hangtani asszimilációk átnyúlnak a szóhatáron:

```
<A><z><a><nny><aa>< ><j><á><ny><á><t><sz><i><ggy>a
```

ahol a <>-ba zárt betűk a kiejtési hangokat reprezentálják, melyeket a lehetséges felszíni karakterek esetleg szóközökkel tűzdelt sorozata helyettesíthet:

```
<nny>=[(ny)|(ny )|([([n|(nn)|(ny)|(nny)|(n )|(nn )|(ny )|(nny )][(ny)|j]]
<ggy>=[(gy)|(gy )|([([d|(dd)|(gy)|(ggy)|(d )|(dd )|(gy )|(ggy )][(gy)|j]]
<aa>=[a|(a a)]
<z>=[z|(z )]
<j>=[j|(j )|(ly)|(ly )]
```

A fenti hangmintának persze négy olyan elemzése van, amit a mondatellenőrző is helyesnek talál:

```
Az anya lyányát szidja
Az anya a lyányát szidja
Az anyja lyányát szidja
Az anyja a lyányát szidja
```

de ez már túlmegy a szóhatáron. Mindenesetre ezt a programot kizárólag a morfológia automatás reprezentációjával sikerült úgy megoldani, hogy a válaszidő hosszabb mondat esetén is másodperceken belüli. A hang-, illetve karakterfelismerő eredményét is DAG-ban reprezentáljuk, és az így kapott két véges automatát, a nyelvi elemzőjét és a bemenetből származót paralel futtatjuk.

4.13.4 Elválasztók

A második tömegfelhasználása a szóelemzőknek a szóelválasztó programok. Ha a sor végére nem fér el a következő szó, akkor ketté lehet törni bizonyos helyeken. Az elválasztásra a magyar nyelvben kemény szabályok vannak. Az alapszabály, hogy összetétel helyén és kötőjelnél lehet elválasztani a szavakat. Ebből a szempontból a prefixek szóösszetételnek számítanak. A többi helyen a szótagolás általános szabálya működik: Két magánhangzó között szótaghatár van. Ha szomszédosak a magánhangzók, akkor egyértelmű a hely. Más esetben egy mássalhangzót kell a második, a többit pedig az első részbe csoportosítani. A többjegyű duplázott mássalhangzók kettévágása úgy történik, hogy a vágás előtti és után duplázás nélküli többjegyű formát kell írni. Egy karakter ugyan alkothat önálló szótagot, de azt leválasztani a szó elején és végén nem szabad, Ez annyiban is igaz, hogy ha összetett szó második tagjának első szótagja csak egy magánhangzó, akkor ezután közvetlen nem lehet a szót elválasztani. Az elválasztott szó első darabján egy kötőjellel jelezzük, hogy a következő sor még hozzá tartozik, kivéve, ha kötőjelnél választunk el. Ilyenkor a kötőjelet nem szabad duplázni.

A magyar elválasztási törvény szigorú, Nem minden nyelven ilyen kemények a szabályok. Angolnál például általában rövidek a szavak, ezért a kiadványok szépsége ritkán sérül, ha egyáltalán nem használunk elválasztást. Tulajdonnevet angolban nem szabad kettétörni. Az elválasztás alaptörvénye ott is az, hogy morfémahatároknál kell, ha van, egyébként kiejtés és szokás szerint. Lengyelben nincs kötelező szabály, csak ajánlás. A *pr, tr, br* csoportokat például nem illik kettétörni, de lehet.

Az angol elválasztást a korábban említett listás módszerrel oldják meg. Miután az összes szóalak technikailag is felsorolható, bejelölhetik, hol lehet a szavakat eltörni. Ki is adtak több helyesírási és elválasztó szótárat. Ha összevetjük a Webster és az Oxford szótárait, mégis eltérések vannak. Eltérések, mert időnként máshogy értelmezik a morfémahatárokat, és a szokás sem törvény. Kinek a pap, kinek a papné.

A lengyel már ragozó nyelv, ezért a szóalakok felsorolása elválasztási pozíciókkal nem jó módszer. A szólista nagyobb lenne, mint ami kezelhető és jól tárolható. Emiatt nem a teljes szóra, hanem karaktersorozat-mintákra adják meg az elválasztási ajánlásokat. A minták néhány karakteresek, és ha a minta megfelel a szó egy részének, akkor a minta mellett megadják, hogyan válasszuk el az adott helyen. Persze a szóösszetétel ott is fontos szabályozó. Egy részlet az IBM által használt mintegy 5000 mintát használó lengyel elválasztó adatbázisból (a nagybetűk előre definiált halmazokat jelentenek):

(b)V=(-b)	(b)V=(-b)	(dl)V=(-dl)	(dz)V=(-dz)
(bl)V=(-bl)	(ch)V=(-ch)	(dŽ)V=(-dŽ)	(dż)V=(-dż)
(bŽ)V=(-bŽ)	(cj)V=(-cj)	(dr)V=(-dr)	(d)V=(-d)
(br)V=(-br)	(cz)V=(-cz)	a(dw)ok=(d-w)	…
(bz)V=(-bz)	(c)V=(-c)	(dw)V=(-dw)	

?(din)to=(din-)	?(do)czła=(do-)	?(do)kl=(do-)	?(do)kł=(do-)
?(diplo)sk=(di-plo-)	?(do)dzw=(do-)	?(do)kr=(do-)	?(do)mk=(do-)
?(do)brn=(do-)	?(do)dzwo=(do-)	?(do)ks=(do-)	?(do)mn=(do-)
?(do)chła=(do-)	?(do)gw=(do-)	?(do)kw=(do-)	?(do)mr=(do-)

A módszer lényege, hogy karaktersorozat-minták felismerése alapján engedélyezik az elválasztás helyét. Ehhez hasonló a szabad szoftver világban élő formalizmus. Az 1970-es években ilyen logikájú elválasztót készített Kardos György a SzTAKI-ban magyar szövegekhez, és a Mac világában is sokáig ilyen módszerrel dolgoztak. A kifinomult mintaillesztési módszerek több menetben döntik el a lehetséges elválasztás helyeit: Egy utólagos menetben tiltják le az egy betű levágását a magyarban. Más metódus, ha a hosszabb találat elvét használják. Ez utóbbi lényege, hogy általános egyszerű mintákkal írják le a gyakori eseteket, de ha egyes helyzetekben másként kell dönteni, akkor az elválasztás nagyobb környezetének mintájával pontosítanak. Így lehet a szóösszetételekből adódó, például *vasszeg* jó kezelését elérni. Ilyen rendszerben, ha a mintában az összes szóalak szerepel, akkor tökéletes elválasztás. Ez azonos a listás megoldással.

Magyarban a mennyiségi okok miatt nem lehet teljes lista alapján dolgozni. A mintaillesztésnek is megvannak a korlátai. A minták száma így is magas, és semmi sem biztosítja, hogy minden esetet leírtunk:

elü=e-lü, elül=el-ül, felüle=fe-lüle, belül=be-lül …

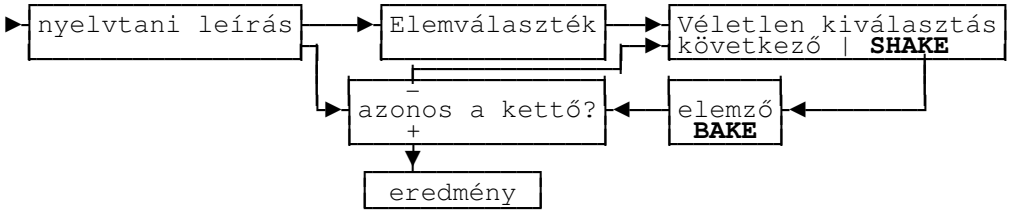
A nagyszámú mintákra két dolog miatt van szükség: a szóösszetételek helyének megállapítása és a betűk felismerése. Az egyszerű mintákból százezrével kéne felvenni, hogy hasonló minőségű elválasztót készítsünk, mint a lengyel vagy az angol esetén. Egy új minta felvétele ellenőrizhetetlen következményekkel járhat. Az igazi megoldás csak szóelemzéssel biztosítható. Ha a minta figyelembe veszi a morfémahatárokat is, akkor párszáz mintával leírható a magyar elválasztás is.

Az elválasztás minőségénél a mellényúlás talán nagyobb hiba, mint a helyesírás-ellenőrzésnél. Ha automatikus elválasztást használnak egy szövegszerkesztőben, akkor – mivel a szöveg módosítása nem a beírás helyénél változtathatja az elválasztást, könnyedén ott marad, és nem javítja ki már senki. Ha a minták mellett a morfémahatárok is felismerhetők – a mintába beépíthetők, akkor Hasonló gond lehet, ha egy szóalakban két magánhangzó között eltér más az elválasztás helye az elemzéstől függően. A *felül* szót automatikusan nem lehet kettészedni a fentiek miatt, mert máshol kell, ha határozó vagy névutó, máshol, ha igekötős ige.

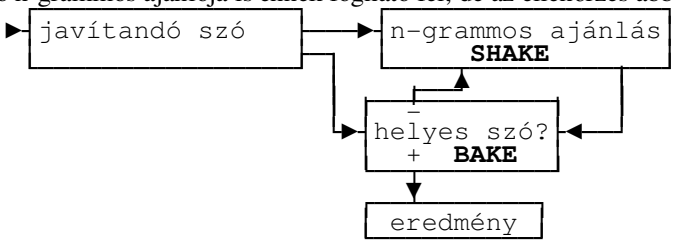
Feladat 1: Gyűjtsön olyan szavakat, melyben elválasztási pozíciók függenek a szó elemzésétől.
Feladat 2: Az ismeretlen szavak elválasztása melyik módszernél nem gond?

4.13.5 Generálás: Shake and Bake a morfológiában

A generálás is fontos feladat. A legtöbb nyelvi modell generatív. Szótannál ez azt jelenti, hogy a formalizmusok azt adják meg, hogy egy szót, ha fel akarunk ruházni nyelvtani tulajdonságokkal, akkor hogyan kapjuk meg a neki megfelelő alakot. Ennek ellenére a generatív leírások és morfológiai implementációk arra vannak kihegyezve, hogyan lehet ennek fordítottját elérni, vagyis a szóformából kell generálni a szó szótári alakját, és a hozzáadott nyelvi információt. Talán a Kimmo és a listás módszer tér el ettől, melyek eleve megfordíthatóak. Ha a konkatenációs módszert fordítjuk meg, akkor az a gond, hogy azt még könnyű megállapítani, hogy milyen nyelvtani kategóriájú morfémák milyen sorrendben követik egymást, de hogy melyik morfémának mely alakját kell, lehet használni, azt nem. Nos, ha az a módszer, hogy kipróbáljuk, melyikkel sikerül megtenni, akkor – ha 4 morfémából áll a szó, és mindegyiknek 4 alakja létezik – akkor 4^4=256 szóalak az esélyes. Ez nem sok, de ha mindegyik szóalakot ellenőrizzük, vagyis a meglevő elemzővel analizáljuk és azonosítjuk az elemzés eredményét a kiindulási információval, már elgondolkoztató, lehet-e valamit spórolni. Generálás esetén a többértékűség a legkevésbé érdekes. Ha egy szóforma megfelel a kívánalmaknak, akkor az jó, és nem kell több. Emiatt a morfémaalakok közt csak addig kell válogatni, amíg meg nem találunk egy kombinációt, ami az ellenőrzés szűrőjén át nem esik. Ezt az eljárást lehet (statisztikailag) gyorsítani. Ha a morfémaalakokat valószínűségük alapján sorba rendezzük, és e sorrend szerint próbáljuk ki, akkor a 256 próba helyett általában 5 próbán belül megtaláljuk az igazit, pontosabban megtalálunk egy helyes szóalakot.



Ezt a módszert, amikor valószínűség alapján választunk, feldobunk egy „jelöltet” majd az így kapott formát elemzéssel ellenőrizzük, nevezik a nyelvtechnológiában **Shake and Bake**-nek. Tulajdonképpen a szövegkorrekció n-grammos ajánlója is ennek fogható fel, de az ellenőrzés abban merül ki, hogy helyes-e a szó:



4.13.6 Indexelők, keresők

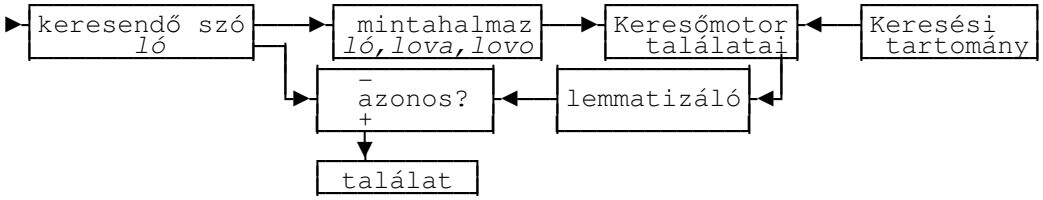
A harmadik tömegfelhasználása a szóelemzőknek a keresőprogramokban rejtőzik. Ha állományainkban keressük, mikor írtunk valamiről, vagy az interneten vagyunk kíváncsiak, hol találunk információt valamiről, akkor keresőprogramokat használunk. A keresők nem arra valók, hogy biztos választ adjanak, hanem arra, hogy segítsenek, mi lehet az az állomány vagy szövegrész, ami valamilyen szempontból érdekes lehet.

A keresés alapja a mintaillesztés. Általában egy meghatározott karaktersorozat a keresendő, de néha bonyolultabb a keresési minta. Vagy egy véges nyelv bármely eleme, vagy egy reguláris kifejezéssel megadott nyelv. Miután mindkét esetben determinisztikus automatára visszavezethető a feladat. Miután a keresendő minta nem szokott bonyolult lenni, feltételezhetjük, hogy belőle a determinisztikus automata, korlátos időn belül előállítható. Emiatt a keresés ideje a keresési állományok hosszával lineárisan arányos. Ez viszont nagyon nagy is lehet. Ha kisebb a keresési tartomány, akkor közvetlen **szabadszövegű keresés**sel megoldható a dolog. Ebben is vannak azért különböző gyorsító eljárások, ami az időigényt akár megtizedeli, de nagy korpuszoknál ez sem segít.

Ha előre ismert a keresés tere – az az állományhalmaz, amelyben keresni akarunk – akkor indexeléssel meggyorsíthatjuk a választ. Ebben az esetben a keresési tartományt előzetesen feldolgozzák, és indexelik szavait, részsztringjeit. Az indexállományba felvesznek minden – a szövegekben előforduló – szót és a szavakban található részsorozatot. A kulcsállomány sokkal kisebb és rendezett, benne gyors a keresés. Ha megadják a keresési szót, akkor elegendő a keresési sztringet kulcsszónak tekinteni, és ennek alapján az indexállomány mutatja meg, hol van találat. Esetleg az index alapján csak azt tároljuk, hogy mely részterület(ek)en, mely állományokban van találat, és utána szabadszövegű kereséssel már futásidőben is gyorsan rá lehet lelni a pontos helyre.

Ha bizonyos értelmes szavakat keresünk, akkor persze nem csak a szó szótári alakja a keresendő. Ha viszont magyarban egy szónak több ezer alakja van, akkor a keresőnek meg kell találnia mindet. Ha *teher* a keresendő, a *teherrel*, *terhet*, *terhetlenül* stb. szavakat is kulcsnak kell tekinteni. Ebből viszont sok van. Szerencsésebb megoldás, az indexelő már eleve a toldalékolt alakokhoz a szó alapalakját rendeli kulcsként. Ha a keresőmotor nyelvfüggetlen, ezt nem lehet megtenni. Az sem megoldás, hogy az összes alakra keresünk, mert a keresőmotorok jó esetben 100 körüli mintát tudnak egyszerre kezelni. Egy érdekes megoldás azokban az esetben, ha a kereső és előindexelő nyelvfüggetlen vagy nincs hozzáférésünk, a következő: A keresendő szónak a tőalakjait (esetleg ennél hosszabb jellegzetes szókezdeményeit) adja meg a keresőmotornak. A keresési processzus eredményeként sok a mellényúlás, de utána a találatokat felülvizsgálhatjuk, és a konkrét esetekben döntünk, kell-e a kereső által talált szó. Így nagyságrenddel kevesebb helyen kell valós elemzést végrehajtani, aminek az ideje esetenként hosszabb, de a kevés eset miatt nem okoz gondot.

Például a *ló* keresésénél a minták a következők: *ló*, *lova*, *lovo*. Ha a szövegben a ló minden lakját megtalálja, de a *lógó*, *lovarda*… szavakat is, amit nem, vagy nem feltétlen keresünk. Ezeket a szó elemzésével dobjuk el.



4.13.7 Intelligens átírások

Szövegszerkesztésnél gyakori gond, ha rájövünk, nem a megfelelő szót használtuk, és helyette egy másikat szeretnénk. Persze létezik egy Find & Replace funkció, de ezzel csak a konkrét szóformát lehet megtalálni, átírni. Most már refrénként írom: a nem vagy alig ragozó nyelveknél ez nem gond, de az erősen ragozó nyelveknél ez nem megoldás. Ha 1990-ben a rendszerváltás miatt a hivatalos iratokban az *elvtárs*, *elvtársnő* szavakat ki akarom cserélni *úr*, *hölgy* szavakra, akkor a sok ezer alakot mind meg kell találni, és egyesével hozzárendelni az *elvtársat*-hoz az *urat*-ot, az *elvtársnővel*-hez a *hölggyel*-t, stb. Ha azonban morfológiai rendszer áll a háttérben, amely képes a szavak elemzésére és generálására, akkor mi sem könnyebb. A feladat megoldását a következő séma adja:

1. Keress a szövegben az átírandó szót. A 4.13.6 alapján ez lehetséges. (pl. *elvtársnővel*)
2. Elemezd meg – a keresés fázisában ez már meg is történhetett. (*elvtársnő[FN]+[INS]*)
3. Az elemzésben írd át a lecserélendő szót az újra. (*hölgy[FN]+[INS]*)
4. Az így kapott nyelvi leírás alapján generáld ki a felszíni formát. (*hölggyel*)

A megoldás bármilyen elemző-generátor eszközzel megoldható. Az egyetlen gond, hogy egy szónak nem feltétlen egy elemzése van, és ha az elemzések közt van olyan, amiben a keresett szó lemmája hiányzik, nem lehet benne biztos a program, hogy cserélni kell. Ha a *vár* szót szeretném *kastély*ra cserélni, akkor még rá lehet jönni, hogy a vár főnévről van szó, máskülönben értelmetlen a feladat, de ha a szövegben *várnak* szerepel, akkor a két elemzés: *vár[FN]+[DAT]* és *vár[IGE]+[t3]*. Ugyan léteznek egyértelműsítő eljárások, de jobb, ha ilyenkor a program interaktív módon rákérdez, kell-e a csere.

4.14 Szótárak

A szótár is gyakorlati problémákat vet fel, de a téma sokrétűsége miatt megér egy külön részt.

A szokásos kétnyelvű szótárak célja, hogy ha nem ismerjük a jelentését egy forrásnyelvi szónak, akkor segítsen abban, mit kell írni helyette a célnyelve, illetve megfordítva, ha értjük a forrásnyelvi szót, de nem ismerjük a célnyelvit, akkor mutassa meg a lehetséges fordítást.

Más szótárak is vannak: képes szótárak, lexikonok, szinonimaszótárak, stb. A fontos az, hogy egy-egy sztringhez hozzárendel egy információt. A sztring többnyire egy szóalak. A hozzárendelt információ a szó jelentéséhez kapcsolódik, és ebből származik a bonyodalom. A jelentést nem tudjuk kódolni, a jelentések között nincs sorba rendezés. Emiatt egy-egy szócikk (egy szóhoz rendelt információ) tagozódik alcikkekre a szó különböző jelentései alapján, tovább jelentésárnyalat szerint, használati környezet szerint, és a szócikkek tartalmazhatnak környezetet és jelentést érzékeltető mintákat, jellegzetes használati példákat is.

A másik gond, hogy egy szónak nagyon sok megjelenési alakja van. A szótár mindet nem sorolhatja fel, ezért egy kanonikus, úgynevezett szótári alakot jelöl csak meg forrásoldalon, és e szerint készülnek a szócikkek. Magyarban a szótári alak névszónál a jelöletlen, ragozatlan alak, igénél a kijelentő mód egyes szám harmadik személyű alak. Németben, oroszban az ige infinitívuszi alakját szokásos szótári alaknak tekinteni, angolban a *to* előtaggal adják meg az ige alakját. Ha a fordítandó szöveget értjük, akkor ez rendben is van, hiszen a folyó szövegben előforduló szóalak szótári alakját könnyű előállítani, de ha számunkra ismeretlen nyelvről fordítunk, akkor bajban vagyunk. Ha ismernénk a szót, akkor tudnánk a szótári alakját is, de ilyenkor minek a szótár? Ha meg a szóalak lényegesen eltér a szótáritól, akkor a szót nem ismerő tehetetlen papírszótár esetén. Ha viszont gépi szótárat használunk, akkor a szóalak kanonizálása, lemmatizálása megoldható.

4.14.1 Papírszótár – gépi szótár

A szótárak hagyományosan sokévi aprólékos munkával jöttek létre. A szótárat most általánosított értelemben használom: legyen az kétnyelvű szószedet, lexikon, telefonkönyv, értelmező, magyarázó szótár. Funkcionálisan a lényege, hogy ha valaminek a jelentését, magyarázatát, leírását keresem, akkor a szó, kifejezés alapján megtaláljam azt az előre megszerkesztett információt, amit gondos munkával a nálam okosabb emberek megadtak.

A papírszótárnak is van struktúrája. Túl azon, hogy elején, végén nyelvtani tábláztok szerepelnek, a lényegi tartalom címszó szerinti ábécébe van rendezve, tele van írásbeli nyelvi **metainformáció**val. Például egy nyelvek közötti szótár azon kívül, hogy a szó más nyelvbeli leírását tartalmazza, egyéb fontos nyelvi, megértést segítő kiegészítéssel tűzdelt. Ezeket betűstílussal jelzik. Ilyenek a vastag-, dőltbetűs jelölések, néha betűzdelt kulcsszavak a metanyelvi információk.

litière *nf* 1. alom 2. *Hist* hordszék ♦ **faire ~ d'une chose** semmibe vesz

litige *nm* 1. jogvita; per: **être en ~** perben áll/pereskedik; **les parties en ~** a peres felek; **point en ~** vitás pont 2. vita; pereskedés: **régler le ~ par voie de négociation** tárgyalásokkal rendezni a vitát; **question en ~** vitás kérdés

A fenti szótárrészletben a dőltbetűs egységek nyelvi minősítések, vastag betűs részek a forrásnyelvi szavak, kifejezések, normál karakterekkel a célnyelvi fordítások jelennek meg. Sorszámmal különítik el a jelentésárnyalatokat, / jellel választja el a belső alternatívákat, ; jel a szinonimákat. : jel vezeti be a mintamondatokat, és ♦ után szerepelnek a sajátos kifejezések, szólások. Szemmel láthatóan szerkezete van egy szócikknek, melyet stílusjegyekkel ellátott részletek teszik olvashatóvá. A szerkezet szótáranként különbözőek. A szerkesztők határozzák meg, mely részt milyen betűtípussal szedjenek. Hagyomány szerint a szótárírók a szövegszerkesztők lehetőségeit kihasználva igyekeztek betartani a szótárszerkesztők előírásait, de az emberi lektoráláson kívül más ellenőrzési lehetőség nem volt. Ha a szótár kellő méretűre dagadt, akkor a hibaszázalék biztosítva volt. Az emberi szem könnyen átsiklik olyan hibákon, melyet automatikusan korrigál az agy, a megértés folyamata.

Már az is nagy haladás volt, amikor nem a nyomdász szedője szedte ki az ólombetűs anyagot. Korábban a próbanyomatok után vastag hibajegyzékek készültek mellékletként, mert a kész oldalak újraszedése sokba került, nem beszélve arról, hogy egy oldal módosítása a többi oldalt is érinthette, ha szövegelcsúszás keletkezett.

а¹ [nrag s] 1. a <hang v betű>; **вимо́вити а** a hangot ejt; **написа́ти а** a betűt ír; 2. felsorolás első pontja; **пункт а** a v első pont; ● **від а до зет** á-tól cettig <elejétől végig>

а² [nrag s] 1. a hang <a C-dúr hangsor 6. hangja>; 2. az a hangot jelölő hangjegy

а³ [ksz] 1. pedig, viszont, ellenben, hanem, de; **він це зробив, а не ти** ezt ő csinálta meg, nem pedig te; **я жарту́ю, а ти се́рдишся** én tréfálok, te pedig haragszol; **я впра́во, а ти вліво** én jobbra, te pedig balra; **не тепер, а тоді́** nem most, hanem akkor; **я́блука ще не дозрі́ли, а вже соло́дкі** az almák még éretlenek, de már édesek; **прийду́ не за́втра, а післяза́втра** nem holnap, hanem holnapután érkezem meg; **не він, а його́ прия́тель** nem ő, hanem a barátja; **не за нього́, а про́ти нього́** nem vele, hanem ellene; **проси́джу годи́ну, а свого́ доб'ю́ся** ülni fogok egy órát, de elérem célomat; **не ті́льки я там був, а й і́нші там були́** nemcsak én voltam ott, hanem a többiek is; **а то** de, különben, azonban; **поспіша́й, а то запі́знишся** siess, különben elkésel; **якби́ сам, а то…** ha még csak maga (lenne) de…; **не ті́льки тому́, що…, а й тому́, що…** nem csak azért, hogy, … de azért is…; **ду́мав і́нших лікува́ти, а дово́диться само́му вмира́ти** máson akar segíteni, de magán sem tud; **стра́шно жи́ти у нево́лі, а ще гі́рше…** szörnyű dolog rabságban élni, de még rosszabb…; 2. és, is, meg; **пі́сля револю́ції, а по́тім пі́сля війни́** a forradalom után és aztán a háborút követően; **він ду́же здібний, а до того́ ще й працьови́тий** ő nagyon tehetséges, és hozzá még munkabíró is; **дити́на, а от розумі́є…** gyermek, és lám érti…; **а са́ме…** azaz, vagyis; 3. **~ ні** ha; **ви му́сите це зроби́ти, а ні…** ezt meg kell csinálnia, ha nem…; 4. [táj] és; **мур між дворо́м а село́м** a birtok és falu közötti fal; **а тако́ж** úgyszintén

а⁴ [msz] 1. no és, nos hát, hát, vajon, talán, bizony; **а де він живе́?** nos, hol lakik?; **а не каза́в я?** hát nem megmondtam?; **а хіба́ ні?** talán nem?; **а чи так воно́?** nos, így van ez?; **а що ка́жеш?** nos, mit szólsz?; **а ході́м вже!** nos, menjünk!; **а розкажі́ть нам, як то було́** nos hát, mesélje el, hogyan is történt?; **а все ж таки́** mindamellett, mégis; **а не ті́льки** hát, nem csak; **а жаль хло́пця** nos, kár a fiúért; **а жаль!** (bizony) kár!; **а втім** v **а проте́** (nos) egyébként, végeredményben, végül is, elvégre; **а зві́сно** ismert, magától értetődő; **а чому́ ж ні, зро́бимо** miért ne, megcsináljuk; **а йді́ть ви всі до чо́рта** hát menjetek a pokolba; **а ти що чита́єш?** hát te mit olvasol?; 2. <kérdés végén> mi? igen? he? hé? ugye?; **піде́мо туди, а?** odamegyünk, mi?; **ти де був, а?** hé, hol voltál?; **дасте́ йому́ дару́нок, а?** ad neki ajándékot, ugye?

а⁵ [isz] 1. <csodálkozás> ó! ah! áh! á!; **а́, прошу́ захо́дьте!** ó, jöjjön be, kérem; 2. <csalódottság> á! ó, jaj! a!; **а ході́м уже́!** a, menjünk már!; 3. <fájdalom> jaj! juj! ó!; **а, боли́ть!** juj, fáj!; 4. **а́-а-а́** a-a-a <gyermek altatását fejezi ki>

Ma papírszótár készítése esetén sem hagyják ki a számítógépet. Nem csak az által, hogy a könnyebb javítási lehetőség miatt szövegszerkesztőket alkalmaznak, hanem azért is, hogy a szótár struktúrája jól követhető, ellenőrizhető legyen. A stílusok, betűtípusok hibáját a szem nem mindig veszi észre, és hiba következtében olyan szócikk is kialakulhat, mely formailag ugyan elfogadható, de a szótárkészítőnek eltérő volt a szándéka. Ráadásul a formázott szöveg az ember számára egyértelműnek tűnik, de a gép – ha gépi használatról van szó – nem tudja eldönteni a struktúra többértelműségét. Például a negyedik jelentés 2. pontjában a *mi? igen? he? hé? ugye?* négy alternatív (szinonim) fordítást jelent, nem egyet, pedig semmi sem szeparálja a részeket. Az ember azonnal jól értelmezi ezeket a formai pongyolaságokat, olyanokat, melyeket algoritmikusan nehéz megfogalmazni. Ellenben ha a szótárat struktúrában szerkesztik, melynek egy megjelenési formája lehet például a papíralapú nyomtatott könyv, ilyen gond nincs. A szerkezetet formálisan adják meg, és a szótárírónak nincs nagyon lehetősége a szerkezetet felrúgni, elvéteni, mert vagy a keretrendszer vezérli a szótárírás folyamatát, vagy

formális szintaktikai eszközökkel felügyelik a helyes struktúra betartását. Ehhez persze olyan formalizmus kell, amely alkalmas a különböző szótárak szerkezetének definiálására – szótáranként változóra, és a formalizmus ne nagyon adjon lehetőséget többértelműségre, félreértelmezésre.

A gyakorlat azt igazolta, hogy a struktúra belső ábrázolására az XML alkalmas. Az XML ugyan ember számára nehezen olvasható, de mindenféle prefix és postfix tulajdonsága miatt könnyű rá elemzőt készíteni, és ezért olyan fordítót, mely az ember számára is tetszetős külsőt adhat a szótárnak. A fenti töredék XML formája a következő:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- XML document processed by MWReflow 1.0 (c) Copyright MorphoLogic, 2008 -->
<!-- XML document created by MarkupWizard 1.1 (c) Copyright MorphoLogic, 1999-2000 -->
<!DOCTYPE dictionary SYSTEM "uhdict.DTD">
<?xml-stylesheet type="text/xsl" href="uh.xsl"?>
<dictionary>
 <entry>
  <head>
   <orth><orth_text>a</orth_text><orth_index>'</orth_index></orth>
   <features><category>nrag s</category></features>
  </head>
  <sub_entry>
   <body>
    <sense_num>1.</sense_num>
    <sense>
     <synonyms>
      <trans>
       <ex_hu_text>a<post_features><comment>hang v betű</comment></post_features>
      </trans>
     </synonyms>
     <example>
      <ex_orth_text>ви́мовити a</ex_orth_text>
      <expression_translations>
       <trans><ex_hu_text>a hangot ejt</ex_hu_text></trans>
      </expression_translations>
     </example>
     <example>
      <ex_orth_text>написа́ти a</ex_orth_text>
      <expression_translations>
       <trans><ex_hu_text>a betűt ír</ex_hu_text></trans>
      </expression_translations>
     </example>
    </sense>
   </body>
   <body>
    <sense_num>2.</sense_num>
    <sense>
     <synonyms>
      <trans>
       <ex_hu_text>felsorolás első pontja</ex_hu_text>
      </trans>
     </synonyms>
     <example>
      <ex_orth_text>пункт a</ex_orth_text>
      <expression_translations>
       <trans><ex_hu_text>a v első pont</ex_hu_text></trans>
      </expression_translations>
     </example>
     <example>
      <proverb_mark>●</proverb_mark>
      <ex_orth_text>від а до зет</ex_orth_text>
       <expression_translations>
        <trans>
         <ex_hu_text>á-tól cettig</ex_hu_text>
         <post_features><comment>elejétől végig</comment></post_features>
        </trans>
       </expression_translations>
      </example>
     </synonyms>
    </sense>
   </body>
  </sub_entry>
 </entry>
 <entry>
  <head>
   <orth><orth_text>a</orth_text><orth_index>²</orth_index></orth>
   <features><category>nrag s</category></features>
  </head>
  <sub_entry>
   <body>
    <sense_num>1.</sense_num>
    <sense>
     <synonyms>
      <trans>
       <ex_hu_text>a hang</ex_hu_text>
       <post_features><comment>a C-dúr hangsor 6 hangja</comment></post_features>
      </trans>
     </synonyms>
    </sense>
```

```
  </body>
  <body>
   <sense_num>2.</sense_num>
   <sense>
    <synonyms>
     <trans>ex_hu_text>az a hangot jelölő hangjegy</ex_hu_text></trans>
    </synonyms>
   </sense>
  </body>
 </sub_entry>
</entry>
<entry>
 <head>
  <orth><orth_text>a</orth_text><orth_index>³</orth_index></orth>
  <features><category>ksz</category></features>
 </head>
 <sub_entry>
  <body>
   <sense_num>1.</sense_num>
   <sense>
    <synonyms>
     <trans>ex_hu_text>pedig</ex_hu_text></trans>
     <trans><ex_hu_text>viszont</ex_hu_text></trans>
     <trans><ex_hu_text>ellenben</ex_hu_text></trans>
     <trans><ex_hu_text>hanem</ex_hu_text></trans>
     <trans><ex_hu_text>de</ex_hu_text></trans>
    </synonyms>
    <example>
     <ex_orth_text>він це зроби́в, а не ти</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>ezt ő csinálta meg, nem pedig te</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>я жарту́ю, а ти се́рдишся</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>én tréfálok, te pedig haragszol</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>я впра́во, а ти вли́во</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>én jobbra, te pedig balra</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>не тепе́р, а тоді́</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nem most, hanem akkor</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>я́блука ще не дозрі́ли, а вже соло́дкі</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>az almák még éretlenek, de már édesek</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>прии́ду не за́втра, а післяза́втра</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nem holnap, hanem holnapután érkezem meg</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>не він, а його́ при́ятель</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nem ő, hanem a barátja</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>не за нього́, а про́ти нього́</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nem vele, hanem ellene</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>проси́джу годи́ну, а свого́ доб'ю́ся</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>ülni fogok egy órát, de elérem célomat</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>не ті́льки я там був, а й і́нші там були́</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nemcsak én voltam ott, hanem a többiek is</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>а то</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>de, különben, azonban</ex_hu_text></trans>
```

```
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>поспіша́й, а то запі́знишся</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>siess, különben elkésel</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>якби́ сам, а то…</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>ha még csak maga (lenne) de…</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>не ті́льки тому́, що…, а й тому́, що…</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>nem csak azért, hogy, … de azért is…</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>ду́мав і́нших лікува́ти, а дово́диться само́му вмира́ти</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>máson akar segíteni, de magán sem tud</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>стра́шно жи́ти у нево́лі, а ще гі́рше…</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>szörnyű dolog rabságban élni, de még rosszabb…</ex_hu_text></trans>
     </expression_translations>
    </example>
   </sense>
  </body>
  <body>
   <sense_num>2.</sense_num>
   <sense>
    <synonyms>
     <trans>ex_hu_text>és</ex_hu_text></trans>
     <trans><ex_hu_text>is</ex_hu_text></trans>
     <trans><ex_hu_text>meg</ex_hu_text></trans>
    </synonyms>
    <example>
     <ex_orth_text>пі́сля револю́ції, а по́тім пі́сля війни́</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>a forradalom után és aztán a háborút követően</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>він ду́же зді́бний, а до то́го ще й працьови́тий</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>ő nagyon tehetséges, és hozzá még munkabíró is</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>дити́на, а от розумі́є…</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>gyermek, és lám érti…</ex_hu_text></trans>
     </expression_translations>
    </example>
    <example>
     <ex_orth_text>а са́ме…</ex_orth_text>
     <expression_translations>
      <trans><ex_hu_text>azaz, vagyis</ex_hu_text></trans>
     </expression_translations>
    </example>
   </sense>
  </body>
…
```

Hát, nem igazán embernek való, szószátyár. A kulcsszavak közt szinte elvész az az információ, amely valóban megjelenik egy nyomtatott szótárban, de jól definiált egymásba ágyazott struktúrával rendelkezik.

A meglévő szótárak többsége nem XML-ben készült. Ha ezeket tovább akarják fejleszteni, akkor érdemes átalakítani olyan formára, melyben a szerkezetet nem külalaki, hanem XML struktúra adja meg. Ehhez meg kell adni, hogy milyen strukturális elemek milyen formai megjelenésben testesülnek meg. A fenti ukrán-magyar szótár eredetileg a Microsoft Word egy korábbi változatában íródott, de kiderült, hogy ez a forma nem felel meg a mai követelményeknek. Az írók különbözőképpen értelmezték a szerkesztők által kiadott formázási előírásait, és nehézkessé vált a szócikkek karbantartása, átrendezése, következetes módosítása. Hogy el ne vesszen az eddigi kódolt anyag, át kellett alakítani a formázásokat struktúrává. XML-esítésének definíciója a következőképpen néz ki:

```
dictionary:          ((entry)?, (~wsp)?, ~endpar)+ ;
entry:               (head, ~wsp?, (body | sub_entry_list | reference) ;
reference:           right_arrow, ~wsp?, orth; // ref_orth ;
head:                (orth | marked_orth), (features | feture_list) ;
orth:                orth_text, orth_index? ;
```

```
marked_orth:~amp, ~wsp?, orth_text, orth_index?, ~wsp?, ~zero ;
features:           ((((~wsp, case | orth_expanded)?, ( ~wsp, paradigm)?, (~wsp, category)?)) | ((
~wsp, paradigm), (~wsp, category), (~wsp, (case | orth_expanded))), (~wsp, head_comment)? ;
feture_list:features, (~wsp?, ~semicolon, ~wsp, features)+ ;
sub_entry_list:     (sub_num)?, features, body, (~wsp?, ~semicolon, ~wsp, (sub_num), features, body)+ ;
body:               (sense | sense_list) ;
sense_list: (sense, (~wsp?, sense_num, ~wsp, (features, ~wsp?), sense)+) | (sense_num, ~wsp?, (features,
~wsp?), sense, (~wsp?, sense_num, ~wsp, (features, ~wsp?), sense)*) ;
sense:              (trans_list, (~wsp?, category)?), (~wsp?, (example)? ;
trans_list: trans, (~colon, ~wsp?, trans)+ ;
trans:              (opt_hu_text, wsp?)?, hu_text, ((wsp?, opt_hu_text)|(wsp, hu_text))*, (wsp?,
opt_hu_text)? ;
example:            ((ex_orth_text, (~wsp?, paradigm)?) | ex_orth_list), ~wsp?, (ex_trans |
ex_trans_list) ;
head1:              (orth | marked_orth), ((((~wsp, case)?, ( ~wsp, paradigm)?, ((~semicolon, ~wsp,
case, ~wsp, paradigm))?,(~wsp, category)?)) | (( ~wsp, paradigm), (~wsp, category), (~wsp, case))),
(~wsp, head_comment)?, (~wsp, orth_expanded)? ;
headfolybef:(orth | marked_orth), ~wsp, paradigm, ~wsp, categfoly, (~wsp)?, (case | alt_orth_text),
(~wsp)?, paradigm, (~wsp)?, categbef, (~wsp, head_comment)? ;
head2:              (orth | marked_orth), (~wsp, case)?, ( ~wsp, paradigm)?, (~wsp, category)?, (~wsp,
head_comment)?, (~wsp, orth_expanded)?, ~colon ;
sub_head:           sub_num (category | ukr_letters)+ ;
body1:              (sense_nc|sense_list|reference|bad_sense_list|sense_ncom) ;
body2:              example, ~wsp?, (~semicolon, ~wsp?, (example_sc|proverb))*, (~semicolon, ~wsp?)? ;
body_short: sense_short ;
sense_list: sense_num, ~wsp?, n_sense, (~wsp?, sense_num, (~wsp?, sub_head)?,~wsp, n_sense)* ;
bad_sense_list:     sense_nc, (~wsp?, sense_num, ~wsp, n_sense)+ ;        // azért "bad_", mert az első
jelentés sorszáma nincs kiírva, de a további jelentéseknek van sorszámuk
trans_syn_list:     (((case | alt_orth_text), ~wsp?) | (paradigm, ~wsp?) | (category, ~wsp?))*, trans,
((~wsp)?, ~comma, ~wsp, trans)* ;
trans_syn_list_nc:  trans_nc, ((~wsp)?, ~comma, ~wsp, trans)* ;
trans_syn_list_ncom: trans_nc, ((~wsp)?, ~comma, ~wsp, trans)+ ;
// trans_list:      trans_syn_list, (~semicolon, ~wsp, trans_syn_list)* ;
trans_list_nc:      trans_syn_list_nc, (~semicolon, ~wsp, trans_syn_list)* ;
trans_list_ncom:    trans_syn_list_ncom ;
sense1:             (trans_list, (~wsp?, category)?), (~wsp?, (example_snc|proverb))?, ((~semicolon?,
~wsp?)?, (example_sc|proverb))*, ~semicolon? ;
sense2:             (example_sc|proverb), (~semicolon, ~wsp?, (example_sc|proverb))*, ~semicolon? ;
sense3:             (trans_list, (~wsp?, category)?), ~semicolon, (~wsp?, example_small), (~wsp?,
example_small)+ ;
sense_short:(trans_short) ;
n_sense:            sense1 | sense3 | sense2 ;
sense_nc:           ((trans_list_nc, (category, ~wsp?)?) | (example|proverb)), ((~semicolon)?, ~wsp?,
(example_sc|proverb))*, ~semicolon? ;
sense_ncom: trans_list_ncom, (~wsp?, category, ~wsp?)?, ~comma, ~wsp?, (example|proverb),
((~semicolon)?, ~wsp?, (example_sc|proverb))*, ~semicolon? ;
trans:              (comment, ~wsp?)?, trans_text, (~wsp?, (comment | bad_comment | category), (~wsp?,
comment2)? )? ;
trans_nc:           trans_text, (~wsp?, (comment | category), (~wsp?, comment2)? )? ;
trans_short:trans_text, ~colon, ~wsp?, col_comment ;
trans_text: ((trans_text_component | (trans_text_component_no_num, wsp?, period+)), wsp?)*,
trans_text_component_no_num, (wsp?, period)* ;
trans_text_component:       (hu_text | opt_hu_text | num ) ;
trans_text_component_no_num: (hu_text | opt_hu_text) ;
proverb:            ~black_circle, ~wsp?, example ;
example:            ((ex_orth_text, (~wsp?, paradigm)?) | ex_orth_list), ~wsp?, (ex_trans |
ex_trans_list) ;
example_snc:(ex_orth_text | ex_orth_list), ~wsp?, (ex_trans | ex_trans_list) ;
example_sc: (((((category)|(comment)), ~wsp?)?, (ex_orth_text | ex_orth_list))), ~wsp?, (ex_trans |
ex_trans_list) ;
example_small:      ex_orth_small, ~wsp?, ex_trans_small ;
ex_orth_list:       ex_orth_text, (~wsp, ~vagy, ~wsp, ex_orth_text)+ ;
ex_trans_list:      (((ex_trans_num, ~wsp?, ex_trans), (~wsp?, ex_trans_num, ~wsp?, ex_trans)+) |
(ex_trans, (~wsp?, ~semicolon, (~wsp?, category)?, ~wsp?, ex_trans)+)) ;
ex_trans:           ex_trans_text, (~wsp?, comment)? ;
ex_trans_text:      ex_trans_text_word, (wsp, ex_trans_text_word)* ;
ex_trans_small:     ((ex_trans_text_component_no_num, period+))+ ;
ex_trans_text_word: (ex_trans_text_component_no_num, (((((comma, ~wsp)?, three_dots)|comma|v),~wsp)? |
mdash | exc | quest | (ex_trans_text_component_no_num)))+, period* ;
ex_trans_text_component_no_num:       (ex_hu_text | opt_hu_text) ;
ex_orth_text:       (ex_orth_text_non_letter_sequence, wsp?)*, ex_orth_text_ukr_letters, ((wsp, (vagy ,
wsp)?)?, (ex_orth_sequence | opt_ukr_text)), (wsp, (ex_orth_sequence | opt_ukr_text))* ;
ex_orth_small:      ex_orth_text_ukr_letters, (wsp?, (ex_orth_text_non_letter_sequence |
ex_orth_text_ukr_letters)), (wsp, (ex_orth_text_non_letter_sequence | ex_orth_text_ukr_letters))* ;
ukr_letters:(ex_orth_text_ukr_letters)+ ;

~endpar:        "[\0160\t \,\;]*\n" ;
~wsp:           "[\0160\t ]+" ;
wsp:            "[\0160\t ]+" ;
orth_text:      "(((\-)|(\.)|(á)|(Á)|['´Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+(
((á)|(Á)|['´Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+)*(\!)?)" bold=1, font=Ukran1;
//  use_orth_text:          "((´)|(á)|(Á)|['Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])((-
)|(´)|(á)|(Á)|['Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])*" ;
alt_orth_text:  "((´)|(á)|(Á)|['Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])((-)|(´)|(á)|(Á)|['Ë-
ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])*[\!\?]?" ;
opt_ukr_text:   "\(((´)|(\!)|(\.)|(\?)|(á)|(Á)|(-)|['0-9Ë-
ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+((,)? ((´)|(\!)|(\.)|(\?)|(á)|(Á)|(-)|['0-9Ë-
ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+)*\)" ;
// use_orth_textp:          "-((´)|(á)|(Á)|['Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])((-
)|(´)|(á)|(Á)|['Ë-ų \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+" ;
orth_index:     "[\185\178\179\8304\8308\8309\8310\8311\8312\8313]+" bold=1, font=Ukran1 ;
~orth_index:    "[\185\178\179\8304\8308\8309\8310\8311\8312\8313]+" bold=1, font=Ukran1 ;
case:                   "-((á)|(Á)|[Ë-ų \469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+(- -((á)|(Á)|[Ë-
ų \469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+)?" ;
paradigm:       "[^\{\}]+" bold=0 ;
categ:          "[^\[\]]*" ;
```

```
categcont:      "foly[^\[\]]*" ;
categperf:      "bef[^\[\]]*" ;
comment_body:   "[^\<\>]+" ;
orth_expand:    "[^\(\)]+" ;
colon:                  ":" ;
~colon:                 ":" ;
~semicolon:     "[;]+" ;
~comma:                 "[,]+" ;
comma:                  "[,]+" ;
~lt: "<" ;
~gt: ">" ;
~left_parant:   "\(" ;
~right_parant:  "\)" ;
~left_bracket:  "\[" ;
~right_bracket: "\]" ;
~left_brace:    "\{" ;
~right_brace:   "\}" ;
~equal:                 "\=" ;
~dash:                  "-" ;
sense_num_empty_nsp:    "(^^^^^^^)?" ;
dash:                   "-" ;
// mdash:               "—";
v:              " v" ;
sub_num:        "(I|II|III|IV|V|VI|VII|VIII|IX|X|XI|XII)\." ;
sense_num:      "[1-9][0-9]?\." ;
sub_num_LETTER: "[A-H][\.\)]" ;
~trans_num:     "[1-9][0-9]?\." ;
~right_arrow:   "→" ;
ex_hu_textquest: "[\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.\,]+([\0160\t ]+[\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.\,]+)*(\?)" ;
ex_hu_textexcl:  "[\-\?'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.\,]+([\0160\t ]+[\-\?'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.\,]+)*(\!)" ;
ex_hu_textexel:  "[\-\?'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.,]+([\0160\t ]+[\-\?'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.,]+)*(\!)" ;
hu_text:        "([\-']+[a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*)|([w-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*[0-9]*)|([a-u][\0160\t ]+[\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|([a-u][\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|(v[\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|([IVX][\-'a-zA-HJ-UWYZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)?)|([A-HJ-UWYZ][\-'a-zAZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)" ;
// hu_textxxx:  "([\-'w-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*)|([a-u][\0160\t ]+[\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|([a-u][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|(v[\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|([A-HJ-UWYZ][\-'a-zAZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)" ;
hu_text_no_dep:  "([\-'w-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*)|([a-u][\0160\t ]+[\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|([a-u][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)|(v[A-ZaeijlnorsuxyzÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*)|([IVX][\-'a-zA-HJ-UWYZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zÁ-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)?)|([A-HJ-UWYZ][\-'a-zAZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+)" ;
ex_hu_one:      "[a-v]" ;
quoted_text:    "[\"][\-'a-zA-HJ-UWYZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêŵ][\-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]*[\"]" ;
transq:                 "[A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ]+\?" ;
//ex_trans_text:    "([0-9][0-9\-\:]* )?([a-u],)?(([\-a-vA-Z]|(IV. Károly))(([\0160\t ]+(([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))*)?(,)?([\0160\t ]*(([\-'w-zA-HJ-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ][\-'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))?(\.\.\.)?)" ;
ex_trans_text:          "([0-9][0-9\-\:]* )?([a-u][\.\,] )?((([\-a-vA-HJ-Z]|(IV. Károly))(((,)?[\0160\t ]+(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))))|(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))|([\-'-A-HJ-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([0-9][0-9\-\:]*[\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))(((([^\(\)A-Z][\-–\.'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))*(([\0160\t ]*(([0-9][0-9\-\:][0-9]+)|[\!\?]|(\.\.\.)))?" ;
ex_trans_text:          "([0-9][0-9\-\:]* )?([a-u][\.\,] )?((([\-a-vA-HJ-Z]|(IV. Károly))(((,)?[\0160\t ]+(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))))|(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))|([\-'-A-HJ-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([0-9][0-9\-\:]*[\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))(((([^\(\)A-Z][\-–\.'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))|((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-HJ-UW-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([A-HJ-Z][\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))|((a )?[0-9][0-9\-\:]* [\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))(((([^\(\)A-Z][\-–\.'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))|((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-HJ-UW-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([A-HJ-Z][\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))|((a )?[0-9][0-9\-\:]* [\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))*)*([\0160\t ]*(([0-9][0-9\-\:][0-9]+)|[\!\?]|(\.\.\.)))?" ;
ex_trans_noel:          "([0-9][0-9\-\:]* )?([a-u][\.] )?(((([\-a-vA-HJ-Z])(((,)?[\0160\t ]+(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))))|(((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])))|((([^\(\)A-Z][\-–'A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*\))|([\"][ \-\?\!'a-zA-ZÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\.]+[\"])|([\-'-A-HJ-UW-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!])|([A-HJ-Z][\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))|((a )?[0-9][0-9\-\:]* [\-'-A-Za-zÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]))*(([\0160\t ]*(([0-9][0-9\-\:][0-9]+)|[\!\?]|(\.\.\.)))?" ;
ex_orth_text_ukr_letters:    "(~)?(~)?((à)|(Á)|[É-ч \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+(~)?";
ex_orth_text_ukr_letts:      "((á)|(Á)|[É-ч \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+";
// ex_orth_sequence:                    "((\.)|(~)|(,)|(')|(\-)|(!)|(–)|[%?0-9àÁ]|[É-
ч \1279\469\470\256\257\482\483\274\275\332\333\362\363\300\301\296\297\298\299\366\367])+" font=Ukran1 ;
ex_orth_sequence:                    "[^\0160\t \n<>\(\)\[\]\];\:\,][^\0160\t \n<>\(\)\[\]\];\:]*" font=Ukran1 ;
opt_hu_text:    "\(((([\-–\.'a-zO-9ÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!][\-–\.'A-Za-zO-9ÁÉÍÓÖŐÚÜŰáéíóöőúüűâêîôûŵŷ\?\!]*)|ъ)|(Angliában)|(Franciaországban[^\)]*)|(IV. Iván)|(Diesel-)|(Korzika és Szardínia szigetén)|(VIT)|(USA)|(Hughes-féle)|(Szűz Mária)|(Mohr-féle))\)" ;
latinb:                 "[A-Z][\-–\.a-zäó, ][\-–\.A-Za-zäó, ;]+" ;
vegyjelb:       "[A-Z][a-z]?" ;
num:            "[0-9]+" ;
numb:                   "[0-9][0-9\-\:][0-9]+" ;
period:                 "\." ;
proverb_mark:   "[•○]" ;
ex_trans_num:   "[a-z]\)" ;
~amp:                   "&" ;
exc:            "!" ;
quest:                  "\?" ;
~zero:                  "0" ;
tilde:                  "~" ;
vagy:                   "v" bold=0 ;
three_dots:     "\.\.\." ;
depend_trans:   "v[km]i(t|nek|ve|re|ről|n|ben|be|ből|hez|től|nél|ért|vé)|vkit, vmit|vmeddig|vhonnan|vhol|vhova|vhogy";
dep_trans:      "v[km]i(t|nek|ve|re|ről|n|ben|be|ből|hez|től|nél|ért|vé)|vkit(,| v) vmit|vmeddig|vhonnan|vhol|vhova|vhogy";
depend_orth:    "(~ )?(((((на(д)?)|пρότи|від|з(а)?|до|у|по|пе ред|із)|((((на(д)?)|пρότи|від|з(а)?|до|у|під|по|пе ред|із)
)?(((ч|к)(ому|омū|orō|ōro|orō|ōro|oro|ōno|им))|бїля|що|кудī)))((,|;| v)? (на |з(а)? |пρότи |від |до |у |під |по |пе ред |із
)?(((ч|к)(ому|омū|orō|ōro|orō|ōro|oro|ōno|им))|бїля|що|кудī))*)|(/+( )?inf))?)" ;
dummy_trans:    "□";
```

A struktúra azért bonyolultabb, mint az korábban jelzett végleges, mert sok helyen csak így lehetett egyértelműsíteni a szövegrészek stílusainak funkcióját. Ebből viszont már könnyen lehetett konvertálni egy olyan egyszerűsített XML állományt, amelynek a struktúrája és eseti megjelenítése egy XSL definíció alapján követhető:

```
<?xml version="1.0" encoding="UTF-8"?>
```

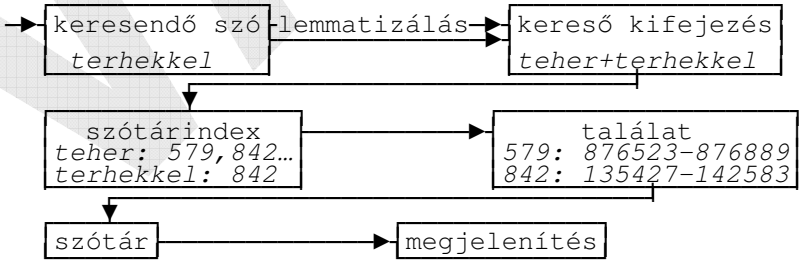
```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output     method="html"
                indent="no"
                encoding="UTF-8"/>
<xsl:strip-space elements="*"/>
<!-- elements having element contetnt -->
<xsl:template match="dictionary">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="entry">  <p>  <font size="3">  <xsl:apply-templates/>  </font>  </p>
</xsl:template>
<xsl:template match="sub_entry">
  <xsl:if test="(preceding-sibling::sub_entry)or(following-sibling::sub_entry)">    <br/>  </xsl:if>
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="post_features">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="body">
  <xsl:if test="preceding-sibling::body">;</xsl:if>
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="head">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="alt_head">  <b>;</b>  <xsl:apply-templates/> </xsl:template>
<xsl:template match="orth">
  <xsl:if test="not(parent::head)"><xsl:text> </xsl:text></xsl:if>
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="features">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="reference">
  <xsl:text> </xsl:text>&#x2192;
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="example">
  <xsl:if test="(preceding-sibling::example)or(preceding-sibling::synonyms)">;</xsl:if>
  <font color="#006600">
  <xsl:apply-templates/>
  </font>
</xsl:template>
<xsl:template match="synonyms">
  <xsl:if test="preceding-sibling::synonyms">;</xsl:if>
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="expression_translations">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="sense">  <xsl:apply-templates/> </xsl:template>
<xsl:template match="trans">
  <xsl:if test="preceding-sibling::trans"><xsl:text> </xsl:text>&#8729;</xsl:if>
  <xsl:apply-templates/>
</xsl:template>

<!-- #PCDATA elements -->
<xsl:template match="proverb_mark">
  <xsl:text> </xsl:text>
  <xsl:apply-templates/>
</xsl:template>
<xsl:template match="orth_text">
  <xsl:if test="not(parent::orth)"><xsl:text> </xsl:text></xsl:if>
  <b>  <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="orth_index">
  <b>  <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="category">  <xsl:text> </xsl:text>[<xsl:apply-templates/>] </xsl:template>
<xsl:template match="case">
  <xsl:text> </xsl:text><b>  <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="orth_alternate">
  <xsl:text> </xsl:text><b>  <xsl:apply-templates/>  </b>
</xsl:template>
```

```
<xsl:template match="paradigma">  <xsl:text> </xsl:text>{<xsl:apply-templates/>} </xsl:template>
<xsl:template match="orth_expand">
  <xsl:text> </xsl:text><b>   <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="comment">  <xsl:text> </xsl:text>(<xsl:apply-templates/>) </xsl:template>
<xsl:template match="use_orth">
  <xsl:text> </xsl:text><b>   <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="dependence">
  <xsl:text> </xsl:text><b>  <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="ref_num">  <xsl:text> </xsl:text><xsl:apply-templates/> </xsl:template>
<xsl:template match="ref_num_end">  â_"<xsl:apply-templates/> </xsl:template>
<xsl:template match="sub_num">   <xsl:text> </xsl:text><xsl:apply-templates/> </xsl:template>
<xsl:template match="sense_num">  <xsl:text> </xsl:text><xsl:apply-templates/> </xsl:template>
<xsl:template match="ex_orth_text">
  <xsl:text> </xsl:text><b>  <xsl:apply-templates/>  </b>
</xsl:template>
<xsl:template match="ex_hu_text">  <xsl:text> </xsl:text><xsl:apply-templates/> </xsl:template>
<xsl:template match="sci_id">  <xsl:text> </xsl:text>(<xsl:apply-templates/>) </xsl:template>
<xsl:template match="trans_index">  <xsl:text> </xsl:text><xsl:apply-templates/> </xsl:template>
<xsl:template match="id"> <!-- internal id, not to be displayed -->
</xsl:template>
</xsl:stylesheet>
```

Az összetettebb szótáraknak ennél sokkal bonyolultabb leírása van. Az akadémiai nagyszótárak például mélyebb beágyazásokat, több alternatívát tartalmaznak. Az XML leírás azt is lehetővé teszi, hogy olyan elemek is bekerüljenek, melyeknek a végső papírformához nincs közük. Szerkesztői megjegyzések, javítás nyomai. Ami még lényegesebb, hogy a szótár elektronikus felhasználását elősegítő információkat is rejthetünk el benne, mint például a keresési algoritmust elősegítő indexkifejezéseket.

Egy papírszótárban az tud igazán keresni, aki a forrásnyelvet jól ismeri. A szótárban levő szöveg nem mindig a szó szótári alakját használja. Ha magyarul nem beszélő azt látja a magyar újságban, hogy *háborús terhekkel sújtott terület*, a teher szót nem találja meg, mert már a harmadik karakternél messze jár az ábécében. Ezért van szükség a keresendő szó lemmatizálására, és csak ez után kell keresni a szótárban. Elektronikusan ez könnyen megoldható. Ráadásul, ha a szótár tételeiben levő szavak akkor is indexelve vannak, ha nem a címszó, akkor meg lehet lelni azokat a kifejezéseket is, melyek esetleg más címszónál szerepelnek, mint a keresendő, de mégis releváns lehet a felhasználónak.

Tehát egyrészt a szótár elemeit, szavait indexelni kell akkor is, ha nem szótári alakban szerepelnek. Ez nem kis munka, de mivel a szótárkészítés fázisához tartozik, előre elkészíthető az indexállomány. Másrészt a bejövő kéréseket konvertálni kell: nem csak a beírt sztringet, hanem a szavak lemmáit is át kell adni a keresőnek. Mivel itt egy-két szó lemmatizálásáról van szó, ezt el lehet futásidőben is végezni.



Feladat 1: Próbáljon egy nyelvpárra egyszerű – párezer szavas szótárat készíteni elérhető szótárak alapján.

4.15 Szószemantika

A jelentéstannal nem óhajtok sokat foglakozni. Az biztos, hogy a szavak jelentésének szerepe van a szótárszerkesztésnél is, de hogy mi a jelentés, azt senki sem definiálja. A matematikai logikusok logikai formulákat próbálnak a szavakhoz rendelni, az ontológusok a jelentéssel nem, de a jelentések hierarchikus strukturált rendezésével törődnek. Tulajdonképpen alkalmazásfüggő, mit nevezzünk szemantikának.

A későbbi fejezetek a mondattannal foglalkoznak, és a szó jelentése, legalábbis egyes jelentéstani tulajdonságok, fontosak a mondatszerkezet szempontjából. Emiatt a szavakhoz rendelt szemantikai jegyeknek

szerepük van a szintaxisban is. Azt is mondhatnám, a határ, hogy milyen tulajdonság nyelvtani, és melyik jelentéstani, hozzáállás kérdése is. Angolban sok melléknév és főnév egyben ige is. Az, hogy éppen egy szó névszói vagy igei szerepet játszik, fontos a szintaxis szempontjából, de mondhatjuk szemantikai kategóriának a szófaji beosztást. Hogy egy ige mozgást, mozgatást jelent-e, vagy egy határozó idő-, hely- vagy módhatározó, lehet szintaktikai és szemantikai jegynek is nevezni.

Ráadásul a magyar szavaknak belső szerkezetük van. Az egyes morfémák módosítják szófaji, szemantikai tulajdonságait a szónak, de a ragok is fontos szerepet játszanak a mondatszerkezet elemzésénél. Ilyen értelemben mindenképpen szükséges, hogy minimális szemantikai tulajdonságokkal foglalkozzunk. Ez semmiképp nem elég ahhoz, hogy valamilyen absztrakt reprezentációt rendeljünk a mondatokhoz, de ahhoz szükséges, hogy elfogadható elemzést, esetleg fordítást állítsunk elő.

4.15.1 Szófajok a mondattan szempontjából

Az egyszerű mondat szintaxisának leírásánál az alapegységek a szavak és az írásjelek. Ezeket nevezhetjük a formális nyelvészet terminológiája szerint a mondattan terminálisainak. Számuk véges – bár egyes modellek szerint még ez sem igaz. Ezek bizonyos szekvenciái helyes mondatot adnak szöveg elemzésénél, mások szintaktikusan helyteleneket. Persze nem csak az a kérdés, hogy mi helyes, hanem minket mindig érdekelni fog, mi a szintaktikus struktúrája a mondatnak. A terminálison kívül szükség van más kategóriákra is. Ezekhez az egyik legfontosabb, a szófaj. A szófaj az angol értelmezés szerint majdnem azonos a mondatrésszel. Ha azt nézzük, egyes szavak kategorizálása segít-e a mondatokban, kifejezésekben szereplő szavak sorrendjének megállapításában, a szerkezet megállapításában, akkor a kategóriák jó szófaji kategóriák a mondattan szempontjából. Ebből a szempontból fontosak az ige és esetragok figyelembevétele, a névutók, a jelek. Ezek a nemzetközi osztályozások szerint nem tartoznak a szófaj-meghatározó kategóriába, de a mondatszerkezet szempontjából számítanak.

A hagyományos (latin alapú) szófaji beosztás: ige, főnév, melléknév, számnév, névelő, névutó, határozó, módosító és kötőszavak. Ez a beosztás nem tökéletes a mondattan szempontjából. Mindegyik kategóriát érdemes tovább bontani, alosztályokat létrehozni. A segédige lényegesen különbözik a többitől, és a létige is másként viselkedik, mint a többi. A mondattan szempontjából a sorszámnév sokkal inkább melléknév, mint számnév. Hogy magyarban milyen kategóriák hasznosak, arra a későbbi fejezetek mutatnak rá.

A magyarban a következő részletesebb szófajokat érdemes megkülönböztetni:

ige:	főige	cselekvő szenvedő műveltető	*húz, tanít, ereszt, növel, megy* *húzódik, tanul, szenved, ered, nő* *húzat, taníttat, növel, származtat, kelt*
	létige		*van, nincs, lesz, létezik*
	segédige	ragozandó hiányosan ragozódó ragozhatatlan	*fog, akar, szeret, megy, tud…* *Kell* *muszáj, tilos*
névszó	tulajdonnév		*Pista, MTA, kedd, nyár*
	köznév		*alma, fürdősapka, nyafogás*
	főnév-melléknév		*iparos, kereskedő, magyar, néger, protestáns*
	melléknév	mely fajta	*háromlábú (szék), nyári (alma), rövidszőrű (vizsla)*
		milyen	*magas, esendő, felhizlal…*
		melyik	*utolsó, szélső, szebbik, legerősebb, negyedik…*
	számnév	határozott	*sok, három, kismillió…*
		határozatlan	*néhány…*
	kvantor		*mind, összes*
	mértékegység		*kiló, liter, kilométer, marok, fő, decibel…*
névutó	alanyesetes		*felett, fölé, között, mellett, alá, közül…*
	birtokviszonyos		*ellenére, folyamán, számára…*
	ragos vonzat		*keresztül, szembe, át…*
Határozók	igemódosítók		*ne, hagy…*
	módosítók		*nem, alig, csupán, alighanem…*
	határozószók		*főleg, rögvest, talán, ma, otthon…*

		általános	*és, s, meg, vagy, illetve, ha*
	kötőszó	mondatrész	*és, s, meg, vagy*
		páros elő-	*vagy…vagy…, akár…akár…, úgy…mint…, sem…sem…*
Egyebek		páros utó-	*…is…is, …sem…sem…, …se…se*
	mondatszó		*hja, nos, hmm…*
	indulatszó		*jaj, óh…*
	igekötő		*be, ki, le, fel, meg, el, haza, agyon…*

Ennél persze lehet durvább és finomabb felosztást is, amitől a szintaxis is durvább, illetve pontosabb lehet. Ha meg tudjuk különböztetni a határozókat aszerint, hogy alap, hely, idő, mód, vagy mennyiség meghatározására szolgálnak, akkor ez is pontosíthatja a szintaxist. Ezen kívül fontos, hogy a határozónak milyen az irányultsága, terjedelme, pontossága, mert ezek is szerepet játszhatnak a mondatelemzésben. A szavak szemantikája ebben segíthet. Egy fizikai tárgy esetén a *-ba*, *-ra*, *-on* alapesetben helyhatározót képez. Ha a főnév cselekedetet, eseményt fejez ki, akkor ugyanezek a toldalékok általában időhatározót állítanak elő. Emiatt érdemes a toldalékokat és a vele rokon névutókat abból a szempontból osztályozni, hogy melyik, milyen jelentésű határozót állíthat elő.

4.15.2 Esetek, ragok, névutók szemantikája

fehervaru rea meneh hodu utu rea

Az esetragok és a névutók a névszóból határozót állítanak elő. A hagyományos nyelvkönyvek nem ezt írják, de a mondatszerkezet szempontjából nem lehet lényeges különbséget tenni a határozói és a ragos szerkezet között. Talán ott van a különbség, hogy egy esetrag különböző típusú határozókat hozhat létre. Ha a szintaxis leírása megkülönbözteti a helyhatározót az időhatározótól, akkor az *ebben az órában érkezett meg*, illetve az *ebben az órában 24 fogaskerék van* más esetragot jelöl. Ha nem tud különbséget tenni a nyelvtan, akkor csak annyit tudunk, hogy valamilyen határozó keletkezik. Szóalaktanilag nincs különbség. A névutó ebből a szempontból hasonló. Egy lényeges különbség, hogy míg a névutót külön írjuk, az esetragot egybe kell írni azzal a szóval, amit megelőz.

Ha a ragok, névutók jelentéstani tulajdonságait is figyelembe veszi a nyelvtan, netán szemantikai reprezentációt akarunk a mondathoz rendelni, akkor érdemes ennek alapján felsorolni, milyen résztulajdonságokat tulajdoníthatunk az esetekhez, névutókhoz.

	→□	■	□→	↔
időtartam	*+ba, +be*	*+ban, +ben*	*+ból, +ből*	*át, keresztül*
időpont	*+ig*	*+kor*	*+tól, +től*	*át, keresztül*
hely pont	*+hoz, +hez, +höz*	*+nál, +nél*	*+tól, +től*	*át, keresztül*
hely felület	*+ra, +re*	*+n, +on, +en, +ön, (+an)*	*+ról, +ről*	*át, keresztül*
hely tartomány	*előtt*	*területén*	*után*	*mentén át; keresztül*
hely vertikális	*felett*	*szintjén*	*alatt*	*át, keresztül*
idő tartam		*belül, bévül*		*kívül, túl*
hely tartomány	*+ba, +be; bele*	*+ban, +ben; belül, bévül*	*+ból, +ből*	
idő tartomány	*előtt*	*alatt; folyamán; +ban, +ben; belül, bévül, alatt*	*után*	*át, keresztül*
hely, birtok	*+nak, +nek*	*+nál, +nél*	*+tól, +től*	
számláló		*+anként, +enként, +onként, +önként*		
eszköz, társ		*+val, +vel; +stul, +stül, +astul, +estül, +ostul, +östül; együtt*		*+tlanul, +tlenül, +talanul, +telenül, +atlanul, +etlenül; nélkül*
mód		*+ként; +ul, +ül +képp, +képpen; +n, +an,+en, +on; +mód, módon*		
réteg	*+rét*			
földrajzi		*+t, +ott, +ött*		
tárgy, mennyiség		*+t, +at,+et, +ot,+öt*		
hely határok	*közé*	*között, közt*	*közül*	
idő határok		*között, közt*		
támogató	*mellé*	*+ért; mellett; érdekében*	*mellől*	*ellen*
hely közel	*+hoz, +hez, +höz; mellé; felé*	*+nál, +nél; ért; mellett*	*+tól, +től; mellől*	*mentén; +szerte*
idő közel	*felé*	*felé, tájban*		*táján*

Persze más, pontosabb beosztás is lehetséges, és még számos névutót vehetünk számba…

4.16 Minőségellenőrzés, A-B teszt

A nyelvi eszközök minősítése fontos feladat. Már korábban tárgyaltuk a lefedettség, a mellényúlás és a pontosság fogalmát. A dolog azon múlik, tudjuk-e mérni ezeket. A gond ott van, hogy a gyakran előforduló esetekben az eszközök jól működnek. A középkategóriában és a kis valószínűségű egyedeknél viszont a kiértékelés teljesen megbízhatatlanná válik. 1993-ban három magyar helyesírás-ellenőrzőt próbáltunk tesztelni. A sebességét, a memóriaigényét pontosan meg lehetett mérni. A minőség viszont nem egy jól rendezett mérce alapján történik. Egyes szövegeknél nem volt lényeges különbség a lefedettségben, vagyis a felismert szóalakok száma a tesztkorpuszon nem volt számottevő. A mellényúlást viszont nem lehetett mérni jól, mert ahhoz tudni kellene, melyek a szöveg helytelen szavai. Már pedig azok ritkán fordultak elő, emiatt egyes szavakra tesztelhettük, de ez nem automatikus teszt. Nem beszélve arról az esetről, amikor a szövegben elírás volt, de az elírt szóalak szabályos szóformát eredményezett.

Minősítést egyébként is csak a felhasználó adhatna, ha lenne alkalma összevetni azonos célt szolgáló eszközöket. A tesztet végzők viszont határozottan érzékelik, melyik ellenőrző a jobb. Ha viszont objektív mérce kell, az legfeljebb nagy minőségi különbség esetén lehetséges.

A megoldás kézenfekvőnek tűnik. Ha nem tudjuk összevetni a gép döntését a valósággal, akkor azt kell megnézni, mikor döntött két módszer (adatbázis) másként, és azt kell vizsgálni, melyik döntött helyesen. Tehát nem lefedettség, mellényúlás és pontosság amit mérünk, hanem azt vizsgáljuk, hogy abban a térben, ahol a két ellenőrző másként döntött, melyiknek volt többször igaza. Ami az egyiknek helyes döntés, a másiknak biztos, hogy a lefedettség vagy a mellényúlás romlására utal. Ha a két ellenőrző 95 százalékos lefedettséget ért el, akkor a szövegnek elméletileg legfeljebb 10 százalékánál kell dönteni emberi kompetenciával, melyiknek volt igaza, de a gyakorlatban ez 1-2 százalék alatt maradt. Hogy aztán hogyan súlyozzuk a különböző hibákat, az már torzíthatja a végső konklúziót. Azzal a ténnyel, hogy a minőség matematikailag nem jól rendezhető tulajdonság, hanem legfeljebb olyan többdimenziós tér, melyben a jóság nem lineáris függvénye a tulajdonságnak, szembe kell nézni.

Akkoriban létezett egy olyan folyóirat – ma már csak webes változata van meg – amely a HiFi készülékekről számolt be. Nos, a szerkesztőknek az volt a véleménye, hogy léteznek ugyan paraméterek, ezek fontosak, de egy szint felett már kevés a teljesítmény, a torzítás és egyéb specifikáció megadása. Ha meg akarom tudni, melyik a jobb készülék, használat közben, vaktesszttel – mindent a fülnek, semmit a szemnek – kell kipróbálni azokat, és szakemberek, vájt fülűek döntsenek, melyiket tartják jobbnak. Nos, az eredmények mindig tanulságosak voltak. Időnként a bírák egységes véleményen voltak, de gyakrabban értékelték a készülékeket különféleképpen. Sőt, szempontok szerint is eltérőek voltak az értékelések. Az egyik hangdoboz dzsessznél mutatkozott jobbnak, a másik klasszikus nagyzenekar hangját adta vissza pontosabban. Szóval a minőség egy szint felett nem egydimenziós érték. Így van ez a nyelvi eszközöknél is. Emberi döntést kell bevonni, de az emberi beavatkozást nem lehet minden szónál igénybe venni. Ott kell, ahol valószínűleg kimutatható minőségi kifogás. Ez pedig az **A-B** vagy **kereszteszt**. Lehet, hogy az egyik ellenőrző irodalmi szövegeken remekel, míg a másik szakszövegekre idomultabb. Az egyik ellenőrzésben pontosabb, a másik javítási ajánlataiban készségesebb.

Ezek a kereszttesztek legtöbbször a lefedettségben mutattak számszerűsíthető különbségeket. A mellényúlás jellegű hibák ritkábban fordultak elő. Ennek a fő oka, hogy nagymennyiségű tesztanyag többnyire csak lektorált, már javított szöveg áll rendelkezésre. A folyóiratok, könyvek nyers gépi formáit a szerzők sem hozzák nyilvánosságra. Ezen is lehet segíteni. két módszer van: Helyes szövegeken statisztikai alapon rontunk, vagyis elütéseket generálunk. A másik lehetőség, hogy a **3.7** fejezetben leírtak alapján generálunk szöveget. Mindkét esetben kellő valószínűséggel alkothatunk olyan szavakat, melyekre a kereszttesztben résztvevő ellenőrzők másként viselkednek.

Feladat 1: Végezzen kereszttesztet azonos nyelv különböző szpellerével.
Feladat 2: A tesztet úgy is végezze el, hogy generált szövegen teszteljen.
Feladat 3: Teszteljen különböző helyesírás-ellenőrzőket, és nézze meg, mekkora a pontossága és a lefedettsége.
Feladat 4: Hasonlítsa össze azonos nyelvre működő ellenőrzők ajánlatát. Mennyit, milyen gyorsan ajánl, és hanyadik ezek közül az adott szövegben kívánatos?

5. Mondattan

Lehet, hogy a magyar nyelv a Jóisten adománya, de az azt leíró nyelvtan biztos emberi csinálmány.
(Farkas Ernő)

A mondat definiálását hasonlóan nem adom meg, akár a szóét. Programozási nyelveknél a különböző egységeket (utasítást, rutint, változót, konstanst…) matematikai precízséggel definiálnak, ezért nincs gond – amit a szintaxis megenged, azt nevezem mondatnak. Az élő, mondjuk az írott nyelvnél is vannak formai követelmények, de ezek nem annyira kötelezők, és nem adnak olyan támpontot, ami alapján a gép tökéletesen lehessen szeparálná a mondatot.

Mint ahogy a szavaknál az írásjelek és a szóközök jó vezérfonalat adnak arra, mit nevezzünk szónak, a mondatnál is vannak olyan ismérvek, melyek jól koordinálnak. Általában abból indulunk ki, hogy valamilyen szövegszerkesztőben ehető formátumban létezik a szöveg, ezért az itt használatos karakterek a mérvadók. A szabályos mondatot általában mondatzáró írásjel zárja, és az első szavát nagybetűvel kell kezdeni. Mondatzáró karakterek a ., !, ?, de lehet más is. Spanyolban például a kérdő és felszólító mondatnak kezdőkaraktere is van, így ezek a mondatok jól zárójelezettek: ¡…!, ¿…? Mivel egy mondatot nem szokás megtörni bekezdéshatárral, ha vége van a bekezdésnek, a mondatnak is vége, akkor is, ha nincs záró karakter. Esetenként a mondat nem kezdődik nagybetűvel, nincs záró írásjel, így nehezebb a felismerés. A helyesen írt szövegben a . lehet mondatvégi, és rövidítés pontja is, vagy akár mindkettő egyszerre, ezért nem egyszerű megállapítani, hol van a mondathatár. Néha helytelen mondatokat is kell elemezni, ezért a feladat még bonyolultabb. Elvileg akkor tudnánk pontosabb eredményt kapni, ha a bizonytalan helyeken megpróbálnánk törni a szöveget, az elemzését így is és törés nélkül is elvégezni, és ha az egyik hibás mondatelemzéshez vezet, akkor a másik lehetőséget fogadjuk el. Ezt a módszert a gyakorlatban nem alkalmazzák, mert – legalábbis magyar nyelv esetében – a kritikus esetek többségében nem segít. Ehelyett inkább némi heurisztika alkalmazásával döntenek ide vagy oda. A mondatelemzés amúgy sem egyszerű feladat, jobb, ha ezt megelőzi egy egyszerűsített mondattagoló algoritmus, ahol legfeljebb a szavak szótani elemzésére támaszkodnak.

Míg a szavak kezelésére bonyolult esetben sem kellett túllépni a reguláris nyelvek osztályán, ezért technikailag egyszerű és gyors eszközök állnak rendelkezésünkre, a mondatok elemzésére néha a környezetfüggetlen nyelvtanok sem bizonyulnak megfelelőnek. Az ugyan igaz, hogy természetes nyelven – például magyarul – korlátozhatnánk a mondatok hosszát pl. 1000 szavasra, ezáltal elvileg a nyelvünk véges, de ez a véges nyelv is oly sok egyedet tartalmaz, hogy felsorolása lehetetlen. Konstrukciójának megadására generatív modelleket érdemes készíteni, ezáltal áttekinthető, kezelhető szintaxis adja meg az elvi helyes mondatok halmazát, mely ezáltal akár végtelen halmaz is lehet.

A nyelvtani modellnek nagy jelentősége van. A nyelvészek által megadott szabályokat formába kell ölteni, máskülönben csak informális, a gép által kezelhetetlen leírása van a nyelvnek. Ha a nyelvészek szabályai nem válnak be a gyakorlatban, át kell azokat fogalmazni. Lényeges, hogy a nyelvtan csak egy modell. A nyelv(használat) ettől eltérhet. Emiatt a kompetencia, vagyis a szabályok által meghatározott nyelv eltérhet a performanciától, vagyis a gyakorlatban beszélt, az emberek által érthető nyelvtől. A jó nyelvi szintaktikai modellek lefedik a nyelvet, vagyis minden helyes mondatot elfogadnak, elemeznek, de persze elfogadnak értelmetlen, vagy ember által nem megemészthető, bonyolult szerkezetű mondatokat is. A performancia viszont elfogadhat helytelen, de az anyanyelvűek által tökéletesen érthető mondatokat. Tehát igazán egyik sem fedi a másikat. A gép szempontjából viszont relatíve kevés információból megadott szabályok kellenek, például formális szintaxis, hogy kezelni tudja a nagyszámú, esetleg végtelen lehetséges mondatot.

Azon nyelvek nyelvtana, melyekben a szavak viszonyát alapvetően a sorrendjük határozza meg, és az izoláló nyelvekhez tartoznak, a környezetfüggetlen nyelvtani leírás megfelelőnek bizonyul. Ha viszont nem így van, akkor más leírási módok kellenének. A bonyolultabb szerkezetű nyelvekre viszont az elemzők sebessége, memóriaigénye válhat túl naggyá. Emiatt más nyelveknél is igyekeznek környezetfüggetlen nyelvtant használni, vagy ahhoz hasonló technikát alkalmazni akkor is, ha ez nem tűnik adekvátnak. Az, hogy a konstrukciós leírások nem maradnak meg a reguláris nyelveknél, attól van, hogy a nyelvészeti modellekben a mondatszerkezetek megadásánál elsősorban a sorrendiség a mérvadó – ettől még lehetne reguláris is a modell –, de a rekurzív egymásba ágyazódást is természetes nyelvi jelenségnek tartják. Ez pedig megköveteli, hogy a leírás formális eszköze legalábbis környezetfüggetlen nyelvtan legyen. Néhány jelenség túlmutat a környezetfüggetlenségen, de a következő Chomsky nehézségi szint, a környezetfüggő nyelvtan már nem igazán praktikus, mert – bár sok minden leírható vele, ami nyelvi jelenség – de a leírás formája nem igazán származik a jelenségből, inkább trükkök azok, melyekkel megoldják a problémákat, s nem maga a probléma

megfogalmazásának nyelve. (Gondoljunk például az ismétlés nyelvének $\{ww \mid w \in \Sigma^*\}$ formális megfogalmazására CS grammatikában.)

Az elemző lényege nem az, hogy segítségével megállapítsuk, mi a helyes és mi a helytelen. Ezt a kérdést megválaszolhatjuk, ha bármilyen formális leírást találunk a nyelvre. Ezzel szemben épp az a feladat, hogy olyan formális leírást adjunk, mely során a mondatok belső szerkezetét feltárhatjuk, tehát ez valóban a nyelvet modellezze. Programnyelvek szintaxisa sem azért készül, hogy megmondjuk, valóban helyesen írtuk le a programot, hanem azért, hogy elemzés, értelmezés után tárgykódot generáljon a fordító.

A szerkezetet az adja, hogy a mondat egyes részei hogyan s miként kapcsolódnak egymáshoz. Ezt a szerkezeti összefüggést, függőséget sok minden jelezheti a felszínen. Egy részük a szavak, kifejezések sorrendjében testesülnek meg, másokat pedig prepozíciók, posztpozíciók, a szavakra tapadó toldalékok jelzik. Nyelvenként más-más a módszer. Míg angolban a mondat alanya megelőzi az igét, a tárgy pedig követi azt, magyarban esetraggal jelöljük az összefüggést. Az biztos, hogy mindkét nyelvben az igéhez tartozik az alany és a tárgy is. A formális leírásban a többi összefüggést is felszíni jelenséggel kell megadni, hogy a gép is képes legyen felismerni a függőségeket. Ezeket a felszíni jelenségeket különböző néven nevezik. Hol egyeztetésnek (pl. a mondat alanya és az igeragnak meg kell egyeznie számban, személyben, szláv nyelvekben a jelző számban, személyben, nemben, esetben megegyezik a jelzett főnévvel, míg a magyarban a jelző megelőzi a jelzett főnevet), máskor vonzatnak (a tárgyas igének kell, hogy legyen tárgya, a tárgyatlannak nem, a *bízik* igének *-ban* esetű kiegészítője van, az *eszik*-nek nincs, viszont lehet tárgya). A nyelvi függőségek tehát hol sorrendi, hol toldalékolási formában öltenek testet. Emiatt azok a nyelvek, ahol nincs morfológia, vagy csak nagyon gyenge, a függőségeket alapvatően a szórend fejezheti ki. A magyarban ellenben az összefüggések nagyobb hányada toldalékokban ölt testet. A szintaxis – tehát a formális leírás – lényege nem az, hogy megállapítsa egy mondatról, helyes-e, vagy sem, hanem az, hogy segítséget nyújtson a szerkezetek felismeréséhez. A programozási nyelvek szintaxisának sem az a lényege, hogy eldöntse, hibás-e a programnak szánt szöveg, hanem az, hogy a felismert programszerkezet alapján futtatható kódot állítson elő. Az élő nyelvbeli mondatok szintaxisa is arra kell, hogy a szöveg megértését, esetleg fordítását könnyítse meg.

Azok az összefüggések, melyeket nem toldalékkal fejezünk ki, magyarban is sorrendiséggel jutnak kifejezésre. Az ilyen kapcsolatos szavak a mi nyelvünkben is egymás mellett szerepelnek. Emiatt érdemes először ezekkel a kisebb egységekkel foglakozni. A magyarban ilyenek a névszói és az igei kifejezések. Hogy hű maradjak az alulról felfelé való építkezéshez, most is ezekkel az egységekkel kezdem.

A további fejezetekben először a magyar nyelv szintaxisát adom meg, ahogyan én látom, majd néhány használható módszerről írok, melyekkel a leírásban szereplő szabályok kezelhetők.

Míg a szavak és a mondatok behatárolására elég pontos robusztus algoritmusok vannak, tehát olyanok, hogy teljes elemzés nélkül is 95% feletti találattal lehet meghatározni, a mondatrészek feltérképezése nehézkesebb. Még egy kifejezésen belül sem lehetünk biztosak, hogy az egymásba ágyazott részeknek hol a határa. Gyakran csak az egész mondat elemzése segít meghatározni, hogy melyik szó melyik részkifejezéshez tartozik.

Nyelvi modell készítésénél elengedhetetlen, hogy a valósággal szembesítsük. Bár én is bízom a nyelvi korpuszokon végzett ellenőrzésekben, statisztikai módszerek használatában, mégis a leghathatósabb módszer, a kipróbálás. Vagyis az, hogy én, mint a magyart anyanyelvi szinten használó, megkísérlek példákat, ellenpéldákat találni a kérdéses modell jóságának ellenőrzésére. A módszer rendkívül hatékony, de egy módosítás mégis szükséges. Ez pedig az, hogy ne egy ember véleménye legyen a mérvadó. Túl sokra nincs szükség, de egy ember ismereti hiányosságát egy-két független kontrol nagy biztonsággal korrigálja. Persze a humán erő mindig drága, de akár az absztrakt elemzés jósága, helyessége, akár egy fordító program pontossága a kérdés, nem lehet objektívebb módszer, mint a szubjektív megítélés – ha az a nyelvet nagymester fokon, anyanyelvi szinten ismeri.

Még egy fontos megjegyzés. A szóelemzésnél is felmerül a kérdés, de ott nem olyan élesen, mint a mondatelemzésnél: minden elemzésnek célja van, és a cél határozza meg, milyen modell alapján érdemes elemezni. Ilyen értelemben nem beszélhetünk arról, hogy a mondat szintaktikus elemzése, logikai reprezentációja, fordítása más nyelvre. Csak arról, hogy a mondat elemzése valamilyen modell alapján, logikai reprezentációja egy konkrét modellben, fordítása egy eszköz segítségével, vagy egy konkrét emberi fordító aznapi hangulatának megfelelően. Így az alábbi leírás nem teljes, lehetne részletesebb, de alkalmazástól függően kielégítő lehet durvább, pontatlanabb leírás is. Én most arról írok, ami a mai számítástechnikai szinten megvalósítható, ezért egy kicsit is komolyabb rendszerben érdemes ilyet vagy ehhez hasonlót alkalmazni.

Feladat 1: Készítsen robusztus mondatokra daraboló programot, amely nem használ szóelemzőt
Feladat 2: Készítsen mondatokra daraboló programot, amely felhasználja a szóelemző eredményét

5.1 A névszói kifejezés

A névszói kifejezés a mondatban jól elkülöníthető rész, mely funkcionálisan valamely mondatrészt tölt be. Ilyen lehet a mondat alanya, tárgya, egy helyhatározói szerepet betöltő rész stb. Ebben az értelemben egy határozószó, vagy egy személyragos névutó is névszói szerkezet: *talán, én, bele, mögöttem.* Ezek külön elemzése nem gond. Annál inkább, a több szavas kifejezéseké. A névszói kifejezés egy esetraggal (esetleg névutóval) végződik, melyet alanyesetű névszók, melléknevek, névelők stb. előznek meg. Itt – az utolsó szót kivéve – nincs rag, amely az összefüggéseket koordinálná, ezért szigorúan sorrendi szabályok lehetnek. Számomra ez a definíciója magyarban a névszói kifejezésnek. A névszói szerkezethez kapcsolódhat még olyan rész, amely nem követel ilyen erős szórendi szabályokat, de ezeket a részek teljesen elszakadhatnak az alább vázolandó szerkezetektől. (lásd: **5.3.2**). Tehát olyan nem igeközpontú szövegrészletet tekintek névszói szerkezetnek, melyben a sorrendiség (többnyire) kötött:

három sárga csőrű kiskacsáról

A fenti példa alapján klasszikusan a következő leírásra lehet jutni:

[Névelő] [Számnév] Melléknév* [Főnév]+jelek+rag

A sorrend kötött, az egyes részek bármelyike elhagyható, hiszen helyes az a kifejezés is, hogy

három sárga csőrűről
három kiskacsáról
háromról

Persze a rag és a jel kell, hogy valamihez tapadjon, ezért valamilyen névszó, kell, hogy maradjon a kifejezésben. Angol nyelven a névszó nem ragozódhat, ezért a főnév mindenképp megmarad a kifejezésben – esetleg egy névmás helyettesíti, de itt is kötött a szórend:

about three ducklings with yellow bills
about three ones with yellow bills
about three ducklings
about three ones

A példákból az is jól látszik, hogy egy egység (*sárga csőrű*) lehet többszavas is. A fenti mondatokban az angol mélyszerkezet nem különbözik lényegesen a magyartól, csak a felszíni. Más a sorrend, és abban tér még el, hogy a főnév nem hagyható el büntetlenül.

Ha pontosabb szintaxist szeretnénk, kiderül, hogy általában a kifejezés eleje is lehet összetettebb, a szótani melléknév helyett a mondattani jelző terminológiát kell használni, számnév helyett mennyiség a megfelelő szó, de a jelzők közt is érdemes különbséget tenni. A fenti leírás nem elegendő, és a kategóriákat, főként a jelző kategóriát pontosítani kell. A további leírásban a **4.15.1.** fejezetben használt szófaji kategóriákat használom.

5.1.1 A névszói kifejezés sorrendi szabályai

A névszói kifejezést sokféleképp lehet megadni. Miután érdemes a modellben az egymásba ágyazást is megengedni, a környezetfüggetlen grammatika lehet a legjobb eszköz. Pontosabban a mostani leírásban bár precedencia jellegű formalizmust is használok, de ez nem azonos a Wirth-Webber féle precedenciával, mert nem szimbólumok közötti szinteket jelentenek, hanem nyelvtani kategóriák (nem terminálisok) előfordulási sorrendiségét akarom kifejezni. A leírás jelen esetben könnyen átírható egy hagyományos környezetfüggetlen nyelvtanra. A szavak szófaji beosztása sem a szokványos, annál részletesebb.

egyszerű névszói kifejezés =
= vonatkozás > névelő > birtokos > szelekció > mennyiség > milyenség > fajta

Az egyes részek opcionálisak, de ha szerepelnek, ebben a sorrendben fordulnak elő. Ha a kifejezés ragos, jeles, akkor ezt csak a legutolsó egység hordozza, kivéve – mint majd látjuk – a vonatkozás rész is viselheti a ragot és a jelet.

A fajta egy objektum általános megadása.

fajta = alfaj > köznév

köznév: tárgyak, fogalmak… általános megnevezése: *kutya, mondat, cinke, só, idea, víz…*

alfaj: a köznévben megadott fogalom pontosítása. Alapvetően két fajtája van. Az egyik, melyet a hagyományos nyelvészeti leírások úgy neveznek, hogy állandósult jelző, mely például a biológiai fajtameghatározásoknál alosztályokat jelöl, de másutt is előfordul: *kékbóbitás rövidlábú cinke, tengeri só, széles szájú*

orrszarvú, de ilyen a *házi feladat* is. Sajnos ezeknek a szavaknak a besorolása függ az őket követő szótól, annak egy kiegészítője. Ha nem olyan szó követi, aminek a pontosítását jelenti, akkor ugyanaz a szó egyszerű jelző. A másik típus emberek kategorizálására szolgál. Ezek a szavak hagyományosan melléknévi és főnévi beosztást is kapnak. Például a foglalkozások, nemzetnevek, ideológiai, vallási, politikai hovatartozást jelentő szavak: *magyar válogatott, fasiszta idea, kommunista mozgalom, lengyel vendégmunkás*… Ez utóbbi esetben a foglalkozásnevet megelőzi a hovatartozást jelentő szó, részkifejezés, sőt ezek közül a nemzetnév szokott az első lenni.

A további jelzők is megkülönböztethetők. Az ismert, hogy egy jelzőre a *milyen, melyik, milyen fajta*… kérdésekkel lehet rákérdezni. De az, hogy melyikkel kérdez(het)ünk, nem mindegy. Az imént tárgyalt részre az utóbbival. A hagyományos *milyen* kérdésre a milyenség rész szolgál. Angolul is különbözik a *which one*, és a *what kind of* kérdésre válaszoló szó szerepe, sorrendje a névszói kifejezésben.

milyenség = melléknév, jelzői szerkezet, melléknévi igenévi szerkezet

Ezzel szemben a szelekcióval nem tulajdonságot akar kifejezni a beszélő, hanem egyed(ek)et akar kijelölni, meghatározni. Ennek, bár szemantikai, illetve pragmatikai kérdésnek tűnik, mégis szerepe van a szintaxisban, ezért érdemes megkülönböztetni a két jelző jellegű részt.

szelekció = kiválasztó jelző > tulajdonnév

melléknév: tulajdonságot, méretet, színt, ízt, stb. kifejező szó: *magas, jó, sokrétű, szerető*…

jelzői szerkezet: többszavas szerkezetek: *tizenhárom emeletes, nagy kaliberű*…

melléknévi igenévi szerkezet: bár lehet egyszerű igenév, mint pl. *játszadozó gyermek*, de lehet több szavas szerkezet. A több szavas szerkezetek egy alfaja a melléknévi igeneves szerkezet. Itt az igenév az utolsó szó, és vonzatai, egyéb határozói megelőzik az igenevet (lásd **5.3.1.**): *tegnap vásárolt, farkas által megharapott, mélyre merülő, Budapest egyik külső kerületében található, üzleteinkben forgamazott*… Az összetettebb jelzők, főként a bővítményes igenevek megelőzik a többit. Talán azért, hogy a mondat esetleges többértelműségét csökkentse a beszélő.

kiválasztó jelző = szelektor, jelzői szerkezet, melléknévi igenévi szerkezet

szelektor: Hagyományosan nem különböztetik meg az egyéb jelzőtől. Kiválasztó pozícióban azonos szerkezetek is lehetnek, mint a milyenségben, viszont a szelektor (sorszámnevek, *-ik* képzős szavak) csak a tulajdonnév előtt, ha szerepel tulajdonnév a névszói kifejezésben. Még egy érdekesség. Ha a névszói kifejezés tartalmaz kiválasztó jelzőt, ést az nem előzi meg más bonyolultabb rész, mindig van vonatkozása, ha más nem, legalább egy névelő. (lásd a következő oldalon)

Azt a szélső zöld üveget hozd fel a pincéből.
A hatodik helyezésért meg kell küzdeni.

A fenti jelző jellegű egységeknek lehet módosítója is. A módosítók szerkezetileg a melléknévhez kapcsolódnak, azt előzik meg. Ilyenek a tagadószó és egyéb néhány felsorolható egyszerű határozószó: *nagyon piros, nem friss, alig meleg, aligha elég*… Hosszabb határozói kifejezések csak igenevet előznek meg, melyek inkább az igének a tartozékai, semmint a belőle képzett melléknévé.

tulajdonnév: A főnevek azon alosztálya, amely funkciója az egyedi azonosítás. Általában nagy kezdőbetűvel írjuk, de tulajdonnévnek tekinthető a tudományos, hivatalos szövegben a képre, bekezdésre, képletre, oldalra stb., utaló szám és betűjel is. Amikor jogi szövegeket elemeztünk, akkor bizonyos köznévnek ismert szavak is tulajdonnévnek számítottak a szövegben: *alperes, vádlott, felperes*…, hisz ezek is azonosították az egyedet a teljes szövegben. Ez persze nagyon különleges eset, és ilyet kizárólag a jogi magyar torz nyelvében láttunk.

mennyiség = számnév > mértékegység

mértékegység: A szokásos mértékegységek, melyek hagyományos köznevek egy része. A mértékegységet időnként alfaj jellegű melléknév is megelőzheti: *három tengeri mérföld*, *öt púpozott evőkanál*, *két öttonnás teherautó ócskavas*. Lehet mértékegységi melléknév vagyis *-nyi* képzővel ellátott köznév, de ebben a formában már jelzőnek számít. A mennyiségnek érdekessége, hogy mértékegység számnév nélkül ritkán fordul elő. Ha viszont pozitív egész számnév szerepel mértékegység nélkül, az általában a *darab* (*fő, példány*) jellegű mértékegységet feltételez. A mértékegységet jelentő szavak többségét fel szokták sorolni. Ha kétségünk van, hogy lehet-e egy szó (alkalmi) mértékegység, tegyük hozzá a *-nyi* képzőt, és ha nem idegen a szóforma, akkor lehet.

számnév: A szokásos számneveket lehet idesorolni. Pontosabban a tőszámneveket. A sorszámnevek a szelektorhoz, a törtszámnevek pedig valahol a számnevek és mértékegységek között találhatók. Nem

csak szófaji szempontból, hanem pozíciójuk is a kettő között helyezkedik el. A számneveknek is lehet módosítója: *körülbelül 100, vagy egy tucat…*

A számnév lehet határozott és határozatlan, tehát nem csak a számok számnevek, hanem a *valahány*, *néhány*, *sok*, *több* szavak is. A számneveket is megelőzhetik számnévi módosítószavak, melyek hasonlóan viselkednek, mint a melléknévi módosítók: *alig, körülbelül, csaknem…*

birtokos = egyszerű névszói kifejezés

A birtokos definíciójánál már rekurzió található. Lényeges, hogy a birtokos számának és személyének meg kell egyeznie a birtok végén található birtokrag számával, személyével, kivéve, ha a birtokost személyes névmással fejezzük ki. Ebben az esetben a többes szám harmadik személyű birtokos csak egyes számban jelöljük, akkor is, ha többes számú. Ha a birtokviszony rekurzív, akkor általában a legbelső (magyarban a legutolsó) birtokos *-nak* ragot kap: *a szomszéd fia jobb lábának nagyujja*. Ettől viszont már lehetőség van arra, hogy a kifejezés kettészakadjon, mert a birtokos ragja és a birtok birtokos jele összekapcsolja a két névszói kifejezést. Emiatt az ilyen birtokviszonyt nem kell feltétlen egy névszói csoportnak tekinteni, hisz nem érvényes a sorrendi kötöttség: *A szomszéd fia jobb lábának bekötötték a nagyujját.* Ha viszont a birtokoslánc nem *nak*-os, akkor csak a legkülső (legelső) birtokos rendelkezhet vonatkozással.

vonatkozás = névelő, utalás > határozott névelő
névelő = határozott névelő, határozatlan névelő
utalás = mutató névmás, visszaható névmás

A határozott és határozatlan névelő a szokásos: *a, az, egy*. A legtöbb nyelvben, ahol létezik határozott névelő, annak eredete a (távolra mutató) névmás, de attól már funkcióban és alakban is eltérhet. A határozatlan névelő a legtöbb nyelven formailag megegyezik az *egy* számnévvel. Így van ez magyarban is. Az érdekes az, hogy a határozatlan névelő kiejtésben szimpla *gy*-vel, míg a számnév nyomatékos *ggy*-vel ejtendő. Ez persze az írott szövegben nem látszik. A határozatlan névelő után csak milyenséget, illetve azt követő részt találhatunk a kifejezésben. Az *egy* szó viszont lehet számnév is, de lehet az *egyik* kiválasztó jelző rövid alakja is:

Iskolába jár az egy, másik bocskort varrni megy…
Péter egy kutyáját láttam kóborolni.

A mutató névmás az *az* és az *ez* névmás ragozott (esetleg névutós) alakja.

Az utalás egy érdekes forma. Egyrészt felveszi a névszói kifejezés esetét, időnként a névutóját (lásd névutók), illetve alanyesetben a birtokjelet is: *Azt a kutyámat. Azé a kutyáé. Magával az ördöggel.* A mutató névmás és a névelő néha egybe olvad. Ilyenkor nem kap esetragot: *E kutyát. Ama lovagokat. Eme házból.* Ha viszont visszaható névmás az utalószó, akkor a névszói kifejezés végén is állhat: *Az ördöggel magával.* Emiatt a fenti félformális leírás nem pontos, de könnyen pontosítható.

A kvantor szintén egy érdekes része a névszói kifejezésnek. A szelektor és a mennyiség közötti forma. Logikából is ismert kavantor a *minden*, de bizonyos prefixszek rendelkező kiválasztó és számnévi névmás is rokonságot mutat a logikában használatos kvantorokkal. A kvantorok önmagukban hordozzák a szelekció és a mennyiség tulajdonságát. Ez határozza meg pozíciójukat a névszói szerkezetben.

Határozott névelővel, vagy birtokossal: *összes, többi.*

Péter összes házi feladatát Feri oldotta meg. A többi szorgalmas diáknak is ötöst adtak.

Határozott névelő nélkül: *mindegyik, bármelyik, akármelyik, valamelyik, semelyik, minden, bármely, akármely, valamely, semely, bárhány, akárhány, valahány, sehány, egy-egy, mindkettő, mindhárom, némi…*

Péter minden házi feladatát elkészítette. Valamelyik szorgalmas diáknak nem adtak ötöst.

Mivel ezek mennyiséget is kifejeznek, nem viselhet többes számot a kvantort tartalmazó szerkezet.
Néhány egyéb névmás is részben kvantorszerűen viselkedik abban az értelemben, hogy előtte már nem sok szerepelhet. A melléknévi, számnévi névmások ezen osztálya nem okoz gondot. Legfeljebb azt a megszorítást jelentik, hogy kategóriájukon belül az elsők: *bármilyen, akármilyen, valamilyen, semmilyen* Például: *Péter három valamilyen utcasarki nőt hozott fel a lakásba*. Ha viszont köznévi névmás szerepel, akkor a szerkezetben utána sem lehet semmi, csak a rag, névutó: *bármi, akármi, valami, semmi,* Habár az a kifejezés helyes, hogy *Péter valami kutyát hozott a lakásba*. de itt a *valami* a *valamilyen* helyett szerepel.

Feladat 1: A mondatban szereplő jelzői szerkezetek melyik kategőriába esnek? Miért? *Újabb devizahiteleseket segítő mentőcsomagon...*

Feladat 2: és ebben a kifejezésben? *Újabb, devizahiteleseket segítő mentőcsomagon...*

5.1.2 Szófajok a mondattan szempontjából

Mint látható, a szótani szófajbeosztás nem feltétlen elég a mondattani elemzés szempontjából. A hagyományos *ige, főnév, melléknév, számnév, névmás, határozó* semmiképpen sem. Ezt kísérlem meg a **4.15.2.** fejezetben. A finomabb beosztás kétrétű. Egyrészt a szótári tételekhez kell párosítani olyan kategóriákat, melyek segítik a szintaxis finomítását, másrészt a képzőknek is megfelelő szófaji beosztást kell tulajdonítani. További felosztást jelent a módosítószavak kategória, melyek hagyományosan határozószóknak vannak titulálva. Ezek közt érdemes megkülönböztetni a melléknévi, számnévi és igei módosítószavakat egymástól.

Szótári munkát jelent az állandósult jelzők jegyzése is, és a névutók kategorizálása (lásd, **4.15.2.**). További kódolást jelentenek azok a vonzatok, melyek nem a szavakhoz, hanem a toldalékokhoz járnak (lásd, **5.3.1.**)

Feladat 1: Hogyan pontosítaná egy hagyományos szóelemzőben a szófaji beosztást a fentiek alapján?

5.1.3 Halmozott mondatrészek

A fenti leírásból még hiányzik valami lényeges. A legtöbb szintaktikai egység lehet halmozott. Ezen azt értem, hogy több azonos kategóriájú egység követheti egymást. Névelőből persze csak egy lehet, de jelzőkből több is. Itt lényeges, hogy az azonos szerepű elemeket, ha nincs mondatrészkötőszó köztük, akkor vesszővel kell elválasztani. Azt, viszont, hogy mi az azonos szerepű, néha csak a szemantika dönti el. A *magas barna lány* kifejezésben a két jelző más (szemantikai) funkciót tölt be, ezért nem kell vesszővel elválasztani, de lehet. Ezzel szemben a *kövér, hájas disznó* kifejezésben logikusabb a vesszővel való elválasztás. Erről az **5.3.1.** fejezetben még szó esik. A halmozás oka sok lehet. Lehet például felsorolás esetleges kötőszóval kiegészítve: *A piros, sárga és fekete tulipán egyaránt szép.* Lehet értelmező jellegű: *Megjött Péter, a kalandvágyó.* A halmozás okának felderítése általában szemantikai ismeretek nélkül nem lehetséges. Részletesebben a kötőszavak használata résznél.

5.1.4 Egymásba ágyazódás a névszói kifejezésekben

A birtokos szerkezetnél már láthattunk egymásba ágyazódást. Ilyen máshol is lehet, hiszen pl. az igenevet megelőzheti az ige vonzata, egyéb kiegészítője, amely gyakran maga is egy névszói kifejezés. Hasonló a helyzet azokkal a kifejezésekkel, amelyek olyan mellékneveket tartalmaznak, melyekhez melléknevet módosító határozói kifejezés tartozik: *esti szürkületben kékes, nappali világosban zöld árnyalatú, minden körülmények között becsületes.* Az ilyen egymásba ágyazódás nem szokott mély lenni, de sajnos gyakran okoz olyan többértelműséget az elemzésnél, melyet csak szemantikai alapon oldhatunk fel. *A Budapesten tavaly forgatott film szereplői által elért siker* esetén nem lehet tudni, hogy a *filmet forgatták*-e *Budapesten*, vagy a *sikert érték el* itt.

5.1.5 A névszói kifejezés szintaxisa

A fentiek alapján formális szintaxist is meg lehet adni a névszói kifejezésre. A gond ott van, hogy a meglevő szótárak nem jelölik a szófajokat oly részletes beosztással, amire a szintaxisnak szüksége van. De ha lenne is, egy egyszerű jelzős főnévnél nehezen lehetne eldönteni, vajon a jelző aktuálisan szelekciót, vagy milyenséget jelent. Emiatt kérdéses, lehet-e, érdemes-e felhasználni a finomabb beosztást. Ha egyszerű szemantikai reprezentációról van szó, például adatbázisban való lejkérdezéshez: *keresem a piros selyem nyakkendő árát*, akkor teljesen mindegy, hogy az áruházban a selyem nyakkendők közül a pirosat keresem, vagy egy piros selyem nyakkendőt. Másrészt a finomabb szófaji beosztás esetenként segít szétválasztani független névszói kifejezéseket. *Melyik szekrényben keressem a nyakkendőt? A zöld felső polcán.* Ez utóbbi esetben finom szintaxis használatával sem egyértelmű, hogy a *zöld* (*szekrény*) birtokosa a *polcnak*, nem pedig színe, de nagyobb valószínűséget lehet adni a birtokviszonynak.

A felvázolt félformális leírás biztos, hogy nem pontos. Néha egyszerűbb nyelvtan is megfelel, máskor összetettebbre is szükség lehet. Ha valaki azt érzi, hogy a beágyazódásokkal csínján kell bánni, akkor akár reguláris kifejezés is adható a névszói kifejezésre. Ekkor persze a belső struktúra elveszhet. Én úgy érzem – és erre nincsenek méréseim – hogy a fenti modell használható, mert a kategorizálás munkája kivitelezhető, ha pedig pontosabbat szeretnénk, akkor kivitelezhetetlenné válik a szükséges adatbázis létrehozása. Nem analitikus módszerrel, pl. statisztikai módszerekkel is eredményre juthatunk.

Egyes névszói csoportoknak pontosabb szintaxist is lehet adni. Ilyen tipikus részszintaxis az időpont megadása:

KORSZAK > ÉV > ÉVSZAK > HÓNAP > NAP > NAPSZAK > ÓRA > PERC …

Például:

Tavaly májusban
Március Idusán
Holnap fél ötkor
Minden reggel öt óra táján

A jellemző, hogy a nagyobb egység után jön a kisebb. A tartalmilag egymásba skatulyázás sorrendi követelménye sajnos nem kötelező szabály. A jellemző sorrendtől el lehet térni, de akkor inkább a halmozott mondatrészekként kell kezelni:

májusban, tavaly
fél ötkor, holnap
öt óra táján, minden reggel

Ha a helyhatározóknak akarunk hasonló pontosított szintaxist, hasonló a jellemző sorrend. A nagyobb egységtől haladunk a kisebb felé:

Magyarország, Budapest, tizenkettedik kerület, Sáfrány utca 7.
Ott a ceruzám, a szobában az asztalon.

Míg az időhatározók aránylag egyszerű módon meghatározhatók, annak megállapítása, hogy mit tekintsünk helyhatározónak, nagyon nehéz. Ha tudjuk, hogy levél címzéséről van szó, persze könnyebb a helyzetünk.
Az angolban például pont fordított a jellemző sorrend, akár időhatározói kifejezésről van szó, akár helyhatározóról, tehát a kisebb egység megelőzi a nagyobbat.

Feladat 1: Készítsen olyan programot, amely egy mondatban kijelöli a lehetséges névszói szerkezeteket.
Feladat 2: Vizsgálja meg, lehet-e reguláris nyelvtant adni a névszói kifejezés behatárolására.

5.2 Az igei kifejezés

Amennyiben a névszói logikára építjük az igei kifejezés fogalmát is, akkor azt kell igei kifejezésnek tekinteni, amit az igével kötött sorrendben írunk. Tehát az igei kifejezés is az egymás után írt elemekből áll, melyeket nem lehet elválasztani az igétől. Az igei kifejezés ebből a szempontból egyszerűbb, mert az ige egyéb tartozékai általában raggal kapcsolódnak az igéhez, ezért akkor is megleljük, ha tőle külön írjuk. Azért igéknél is van két elem, amely nem mehet máshová. Az egyik az igei módosító, amely megelőzi a ragozott igét: *nagyon szereti, ne tedd, nem élhetek, alighanem szereti.* A másik, amit én igeutónak szoktam nevezni, a *volna* szó, amely a múlt idejű feltételes mód esetén jelenik meg, mivel a magyar ragozásban az igeragok e két módját nem viselheti egy szó (régies beszédben a *volt* is lehet, mint a régmúlt – múlt a múltban kifejezője) *tudtam volna, nem kétkedett volna, mondta volt*… Az ige egyéb kapcsolatai nem kötött szórendűek, bár a segédigés kifejezések elemei, az igekötők leggyakrabban közvetlenül, egy sorozatban szerepelnek a mondaton belül. Az igekötő persze, ha közvetlen megelőzi az igét, része a kifejezésnek, talán akkor is hozzátartozik, ha az igekötő és az ige között csak módosítószó van: *ki ne kezdj, fel sem álltam*… Ha követi, akkor már bármilyen más mondatrész is közé jöhet, mint angolban a phrasal verb-ök kiegészítő tagja és az ige közé: *Menj a szobába be. Péter árulta a titkot el. Take your coat off!* Logikailag természetesen az igekötő ilyenkor is az ige része, de az elemzésnek abba a logikájába, hogy az folytonosan írandó részeket próbáljuk szeparálni, abba nem. Hogy hogyan találnak egymásra a külön írt részek, arról a mondatok szintaxisa szól.
A fenti példákból is látszik, hogy érdemes az igei módosítót is az igei kifejezés részének tekinteni.
Kérdés, hogy a segédigés szerkezeteket egy egységnek tekintsük, vagy sem. Miután a segédigés szerkezetben a főige infinitívusszal jár, ezért *nem kell* mindenképpen a segédige mellett *állnia*. Ebből következően az én modellemben külön kezelhető résznek tekintem, de akinek nem tetszik, veheti egybe is. Ez utóbbi esetben viszont bonyolultabb algoritmussal kell elemezni a mondatot.
Feladat 1: Készítse el az igei kifejezés környezetfüggetlen szintaxisát
Feladat 2: Vizsgálja meg, lehet-e reguláris nyelvtant adni az igei kifejezésre.

5.3 Egyszerű mondatok – szabad a szórend

A mondatok lehetnek egyszerűek és összetettek. Bár egy mondatot egy folytonos karakterlánc reprezentál, összetett mondatnál az egyik tagmondat beágyazódhat, így esetleg „kettévághatja" a másikat. De erről az összetett mondatoknál írok. Először a nem összetett mondatok szintaxisát igyekszem megadni. Az sem egyszerű.

A szavak és a mondatok behatárolására elég pontos robusztus algoritmusok vannak, tehát olyanok, hogy teljes elemzés nélkül is 95% feletti találattal lehet meghatározni,

Az egyszerű mondat szintaxisának leírásánál az alapegységek a szavak és az írásjelek. Ezeket nevezhetjük a formális nyelvészet terminológiája szerint a mondattan terminálisainak. Számuk véges – bár egyes modellek szerint még ez sem igaz. Ezek bizonyos szekvenciái helyes mondatot adnak szöveg elemzésénél, mások szintaktikusan helyteleneket. Persze nem csak az a kérdés, hogy mi helyes, hanem minket mindig érdekelni fog, mi a szintaktikus struktúrája a mondatnak. A terminálison kívül szükség van más nyelvtani kategóriákra is. Az egyik legfontosabb, a szófaj, és mint korábban láttuk, a névszói és igei kifejezés.

A magyar nyelv úgy ismert, hogy szabad szórendű nyelvtannal rendelkezik. Valóban, a következő háromszavas mondat szavait tetszőleges sorrendbe tehetjük, a jelentés lényegében nem változik:

Péter szereti Katit. *Péter Katit szereti.* *Katit Péter szereti.*
Katit szereti Péter. *Szereti Katit Péter.* *Szereti Péter Katit.*

Ez a három szóból álló mondat a szavak tetszőleges permutációját felveheti, az alapjelentés mit sem változik. Ez azt is jelenti, hogy a környezetfüggetlen és hozzá hasonló leírási módszerek nem igazán hatékonyak abban az értelemben, hogy a mondatrészek viszonya, egymáshoz rendeltsége a sorrendtől (majdnem) független, azt gyakran esetragok, névutók fejezik ki. A nyelvészeti leírásnak pedig nem az a fő feladata, hogy megállapítsa, hogyan néz ki a helyes mondat, hanem az, hogy a részek közötti viszonyt adja ki eredményül. Ez a célja a kategoriális, az unifikációs, a lexikalista leírásoknak. A cél tehát az, hogy a mondatnak olyan szerkezetét adjuk meg, mely – esetleg nyelvfüggetlenül – a benne levő elemek viszonyát tükrözze egy fában. Ezeknek az irányzatoknak nem elhanyagolható részében a viszonyt függőségnek nevezik. Az elemek, legyenek azok egyszerűek vagy összetettek, függenek egy másiktól. A függőség hierarchikus, és a legtöbb leírás szerint minden rész csak egy funkciót tölt be, tehát egy függő rész csak egy helyre kötődhet. Ez a függőségi modell a magyarban is bevált, de fontos tudni, hogy míg az angolban a függőségi viszony többségét sorrendi szabályok határozzák meg, addig a magyarban a viszony szavakban, ragokban, néha szemantikai tulajdonságokban testesül.

Hogy mit tekintsünk vonzatnak, akár szabad megegyezés kérdése is lehet. Az indoeurópai nyelvek modellezésének gyakorlatában a sorrendi kötöttséget is annak szokták nevezni, sőt formálisan az egyeztetéseket is oda vezetik vissza (alany-igerag, melléknév-főnév esete, számnév- többes szám…), de én az ilyen jellegű szabályokat egyeztetésnek nevezném, a korábban említett sorrendi szabályoknál maradnék a szintaktikus szabálynál, míg az egyéb kötődéseket nevezném vonzatnak.

A vonzat tehát egy fontos fogalom. Legtöbbször az igének a vonzatait írják le. Az angol leírásában az ige vonzata a tárgya, esetlegesen a második tárgy (ami azt fejezi ki, hogy kinek a részére). Többet a formális leírásnál nem szoktak használni, mert ezek azok a részek, melyek alapvetően meghatározzák a mondatrészek sorrendjét, legalábbis, ha a névszói kifejezések jól elemezhetők. A magyarban is sokáig csak azt nézték, hogy az ige tárgyas, vagy nem. Ha a magyarban minden mondatnak lenne is alanya, a tárgyas igének tárgya, akkor itt talán elegendő lenne az ilyen jellegű elemzés.

5.3.1 Vonzatok, vonzatkeretek

A vonzatok fontosságát már említettem, dehogy mi az, azt nem definiáltam. Tulajdonképpen a mondatszerkezet egy olyan kötőeleme, amely egyes részekhez más részeket kapcsol alárendelt módon. Ha az angol nyelvet nézzük, akkor az igéket két fő csoportba osztják. Az egyik a tranzitív, a másik az intranzitív igék csoportja. Amikor mondatot elemeznek, akkor a tranzitív igéknek kell, hogy legyen tárgya, míg az intranzitív igéknek nem. Ez a mondatelemzés szempontjából elegendő, hiszen az alany megelőzi az igét, a tárgy követi, de ha van még egy prefix nélküli névszói kifejezés az ige után, akkor az a másodlagos tárgy.

I read a book.
I tell you a story

Logikailag persze a másodlagos tárgy a magyar részes esetűnek, az elsődleges meg a tárgyesetűnek felel meg. A további mondatrészek vagy határozószók, vagy valamilyen prefixszel kezdődő névszói kifejezésekként jelennek meg a mondatban. Emiatt azt, hogy helyes-e a mondat, egy egyszerű szintaxissal le lehet írni. Ha viszont a szintaxist arra akarjuk használni, hogy fordítsunk, vagy szemantikai reprezentációt készítsünk, akkor ez kevés. A szintaktikai struktúra ilyenkor nagyon lapos, és nem feltétlen derül ki, mely mondatrész mihez tartozik. A szintaxis feladata pedig nem csupán az, hogy eldöntsük, helyes-e egy mondat, hanem az is, hogy a szintaxis által alkotott fastruktúra segítsen a mondat mélystruktúrájának előállításához.

Magyar nyelvű szóosztályozásnál is jelölik, tárgyas vagy tárgyatlan-e az ige. Az igének viszont nem ez az egyetlen paramétere. A *fúl* igének ugyan nincsen tárgya, de paramétere, hogy *mibe fúl valaki*. A *bízik* igével más szintaktikájú mondatot kell építeni, ha *valakire bíz valamit* (tárgyas eset), vagy ha *bízik valakiben valaki*

(tárgyatlan eset). Megint más a szerkezet, ha *megbíz valakit valamivel*. Nos, ebben az utolsó két esetben nem tárgyesetű vonzat is van. Ezen részek felismerése nélkül nem lehet olyan elemzést adni, ami segíthet az értelmezésben, fordításban. Ráadásul a magyarban (is) vannak olyan igék, melyeknek alanyuk sincs. A *villámlik*, *mennydörög, fagy, tavaszodik, havazik* igéknek erőltetett alanyt keresni. Az indoeurópai nyelvekben a nem létező alanyt általános alannyal pótolják, ne hiányozzék valami a mondat szerkezetéből: *es blitz, it is thundering, on gèle*.

Az ige vonzata valami olyasmi, mint a függvény paramétere. Ha az ige felel meg a függvénynek, akkor keresni kell a mondatban a függvény paramétereit, amelyet be lehet helyettesíteni. Vagyis a függvény változóiba kell behelyettesíteni az aktuális megtalált értékeket. Ezt nevezzük a vonzatok lekötésének. Akkor jó a mondat kiértékelése, ha a vonzatokat jól tudjuk lekötni, vagyis megtaláljuk azokat a részeket, melyek kielégítik a különböző vonzatot kielégítő követelményeit.

Az igék vonzatai függnek az ige jelentésétől. Egy jelentés nem egy vonzatot követel, hanem egy vonzatkeretet, amely több vonzatot jelent (mint függvény, több paramétere lehet). Az *ad* igének a tárgya és a részese is paraméter, de az én modellemben az alany is az. Legalábbis, ha a jelentése *átnyújt*, *adományoz*. Az *ad neki* kifejezésben (*beüt*, *megver*) viszont csak alanya és formálisan részes esetű névszói kifejezés a két leköthető változó, tárgya nincs. Az *ad a jó modorra* kifejezésben viszont *-ra* ragú névszói kifejezés az alanyon kívüli vonzat. Ebben az értelemben az ad ige formális leírásában más jelentéshez esetén más vonzatkeret tartozik. Az adott példában a következők:

1. *valaki valamit valakinek*	*A vállalkozó pénzt ad a dolgozóknak.*
2. *valaki valakinek*	*Péter adott neki.*
3. *valaki valamire*	*Ad a véleményére.*

Itt a vonzatkeretek 2-3 eleműek. Formailag ragokkal jelölt névszói kifejezések tölthetik be az egyes vonzatokat. Ha fel tudjuk használni, akkor az is információ, hogy személyt vagy nem személyt jelöl az egyes vonzat. A vonzatok között vannak úgynevezett kötelezőek és opcionálisak. A kötelezőek – még ha hiányoznak is a mondatból – jelenlétük nem maradhat el, illetve, ha elmarad, hiányos a mondat. Az opcionális viszont olyan, amely nélkül teljes értékű a mondat, de ha van, akkor kiegészítő paramétere az igének, mint függvénynek. A korábbi példánkban az első vonzatkeretnek lehet valamire vonzata is:

valaki valamit valakinek [*valamire*] *Péter pénzt adott a koldusnak kenyérre*.

A fenti vonzatkeret alapján, ha tárgy is szerepel a mondatban, az csak az első vonzatkeret lehet, mert a többiben nincs tárgyi vonzat.

Vannak persze olyan vonzatok, melyeket nem ragok (névutók), hanem általánosabb dolgokat jelölnek. A mozgást, mozgatást jelentő igéknél például a honnan és a hova kérdésre válaszoló névszói kifejezések bármelyike lehet paramétere az igének, többnyire nem kötelezően. A vonzat ilyen értelemben szemantikai jegyeket is meghatározhat. A *tölt* igének egyik jelentése (*spend*) például olyan tárgyragos vonzattal rendelkezik – *nem sok időt tölt tanulással* –, ami csak idő(tartamo)t jelenthet, szemben a *vizet tölt a kancsóba* jelentéssel, ahol a tárgy anyagot jelenthet csak.

Az is lehetséges, hogy a vonzat nem általánosabb, hanem speciálisabb. Ily módon a speciális szókapcsolatok is kezelhetők vonzatként. Például a *felrúg* igének a *bakancs* az alannyal egyeztetett birtokraggal és tárgyesetben egy olyan vonzata, mely a mondat jelentését lényegesen megváltoztatja. Míg abban a mondatban, hogy *felrúgta a bakancsát a tetőre* létezik a mondatban olyan rész, amely a HOVA kérdésre válaszol, így a rúgásnak a mozgatás jellegű jelentését lehet csak figyelembe venni, mert nem fér bele a másik jelentésbe ez a mondatrész. Van persze olyan eset, amikor az ige jelentését, vonzatkeretét módosító szóban semmilyen egyeztetésre nincs szükség. Máskor viszont csak az egyeztetett tulajdonságokkal tölthetjük ki a vonzatkeretet. Esetleg a kérdéses névszói szerkezet főneve a vonzat, de a névszói szerkezet egyéb részei szabadon kiegészíthetők.

Feladom magamat.
Beadja a kulcsot
Felverte az arany napi árát

Az ilyen kulcsszavas vonzatokat a nyelvészek egységes kifejezésnek tekintik. Elemzés szempontjából viszont jobb szétválasztani a vonzó és a vonzott részt, már csak azért is, mert ezek a részek a magyarban nem feltétlen egymás mellett jelennek meg a mondatban. Az első példamondat különösen érdekes. Az az érdekessége, hogy a magam visszaható névmás a tárgyeset ragján kívül megkapja az alanynak megfelelő birtokos jelet is, tehát egyeztetni kell az igeragot a visszaható névmás személyragjával.

Feladom magamat.

Feladod magadat.
Feladja magát.

Ilyen birtokragos eset nem csak a visszaható névmás használatánál van:

Eszét veszti
Nem leli kedvét benne.

Néha nem az alany, hanem más mondatrész szerinti birtokragot kell használni:

Kiforgatta az áldozatokat minden vagyonukból.

Más esetben a rag és a birtokrag (valamilyen birtokviszony) kötelező, de nem egyezik meg a birtokrag sem az alannyal, sem a tárggyal, sem más mondatrészével:

Megbízik a szavamban.

A fenti példák közül a visszaható névmásoson el lehet gondolkozni, vonzatnak kell-e tekinteni, vagy a visszaható névmás az igei kifejezés része. Például a latin nyelvek esetén ez praktikusabb megoldás. Mivel a magyarban a visszaható névmásos igék a névmástól elszakadhatnak, ezért talán előnyösebb az ige vonzataként kezelni. Hasonló gond merül fel az igekötők esetében. Mivel az igekötőt az esetek nagy hányadában egybeírjuk az igével, ezért érdemesebb az igei kifejezés szerves részeként tekinteni, akkor is, ha nem írjuk egybe a két szót. Technikailag ez gondot okozhat, de nem komolyat. Ha az igekötős igéknek duplázva vesszük fel a vonzatkeretét, igekötősként, és igekötő nélküli igeként, így formálisan megoldhatjuk a problémát: *ellátja a baját.*

ellát [*baj*+ACC]
lát [*el, baj*+ACC]

Mivel a vonzatok a szó jelentéséhez köthetők, ezért használatuk a lexikalista irányvonalnak megfelelő eszközzel kezelhetők legjobban. Egyes igecsoportok – például a mozgást jelentő igéknek – lehet a vonzatkerete azonos, de általában a konkrét szavakhoz rendelt vonzatkeretek nem sok hasonlóságot mutatnak.

5.3.2 Nem igei vonzatok

Nem csak igéknek lehetnek vonzatai. Ha vonzatnak nevezem azon függőségi viszonyt, mellyel egy mondatrész egy másik szóhoz kapcsolódik, akkor főneveknek, mellékneveknek is lehet vonzata. Tipikus nem kötelező vonzata az *út* szónak, az a mondatrész, ami azt fejezi ki, hogy honnan, hova. Egy embernek, hogy honnan jött, honnan származik. Ezek a vonzatok sohasem kötelezőek.

A kirándulás a hegyekbe jól esett.
A túra Szentendrére elfárasztott.
Péter Szolnokról nem tud síelni.

A fenti kiegészítőket melléknévi igenév segítségével elhelyezhetnénk a névszói szerkezetbe is. Ha beágyazott szerkezetben szerepelne ilyen jellegű vonzat, közvetlenül csak úgy előzheti meg a vonzat a kérdéses szót, ha igenév segítségével kapcsolódik. Ebben az esetben az igenév nélküli vonzat elvész.

A hegyekbe vezető kirándulás jól esett.
A Szentendrére vivő túra elfárasztott.
Szolnokról származó (érkező) Péter nem tud síelni.

Vonzatos melléknévre példa a *büszke* szó mire vonzata: *A fiára büszke anya*
A névutós szerkezetek is felfoghatók formálisan vonzatos szerkezetnek. A (melléknévi) névutó vonzata egy, a névutó vonzatának megfelelő névszói szerkezet. Ha ez a vonzat alanyesetű, akkor a vonzatot megtestesítő kifejezés kötelezően megelőzi a (melléknévi) névutót. Ha nem, akkor a rag alapján megtalálható a vonzatot kitöltő rész. Míg a korábbi névszói vonzatok nem kötelező jellegűek, a névutók vonzatai kötelezők. Elhagyásuk erős hiányérzetet von maga után, főleg az alanyeseteseké.

Közvetlen óra előtt készítették el a leckét.
Szemben a nagy teátrummal megittam egy teát rummal.

A melléknévi névutók vonzata mindig megelőzi a névutót, míg az egyéb ragos névutó vonzata követheti a névutót, sőt, időnként elszakadhatnak egymástól

Bizonyos esetekben nem a szótő maga, hanem egyes toldalékok következménye, a vonzata van. Néhány tipikus példa: *-ú, -ű* képző, ha nem összetett szó, kötelező melléknévi, esetleg mennyiségi, vonzata van, mely megelőzi a képzett szót. A vonzat kötelező, bár felfogható úgy is – nyelvészek ezt teszik – hogy a melléknéves szerkezet, és nem a főnév kapja az említett képzőt:

Nagy hasú török basa.
Szép kiejtésű szónok

Öt hektár területű.
Tizenkét fejű sárkány.

A középfokú melléknév nem kötelező vonzatkerete két elemből áll: *-nál* és *-val* ragot hordozó névszói kifejezés. Az első az összehasonlítás másik felét jelöli, a második pedig a különbség mértékét.

Az Eiffel toronynál sokkal magasabb épület is létezik Párizsban.

A középfoknak létezik másik vonzatkerete is, és a két vonzatkeret közül csak az egyik szerepelhet aktuálisan. Ez a *mint* kötőszavas mellékmondat. Igen, mellékmondat is lehet vonzat, de erről az összetett mondatoknál írok:

Jóval többet költöttek el, mint amennyit a gazdasági helyzetük megengedett.
Az Eiffel torony nem sokkal magasabb, mint a többi Párizsi felhőkarcoló

Vigyázat, felsőfoknál ezek a vonzatok megsemmisülnek, legfeljebb egy *között, közül* névutós nem kötelező vonzatot lehet értelmezni!

Az osztálytársak között legmagasabb diák állt a sor elején.

Ha a névutóhoz kapcsolódó névszói kifejezést is vonzatként kezeljük, akkor a melléknévi névutók annyiban térnek csak el, hogy a melléknévi névutókhoz tartozó névszói kifejezéstől nem válhat el akkor sem, ha a vonzat nem alanyesetű. Ráadásul a kötelező vonzat nem maradhat el a kifejezésből sohasem.

Az állításával szembeni tényeket nem vette figyelembe.

Az igenevek kötelező vonzatai is elhagyhatók nem kötelezők, de ha a melléknévi igenév egy főnév jelzője, akkor vonzatai közvetlenül megelőzik az igenevet:

A Dunán átívelő hidak szépek.
Az átívelő hidak szépek.

Egyes mellékneveknek, melyek segédigei funkcióban is lehetnek, infinitívuszi vonzatuk is lehet. Az ilyen melléknevek egy része valóban segédigéből képződik, de egyéb jelzőknek is lehet hasonló tulajdonsága:

A szűnni nem akaró taps kihallatszott a színházból.
Az iskolába menni vágyó óvodás nem tudja, mi vár rá.
A megenni nyers gyümölcs kemény volt.

Ezek a vonzatok – bár a kapcsolat jelezve van az infinitívuszi végződéssel, mindig megelőzik a jelzőt a névszói szerkezeten belül.

A mellékmondat jellegű vonzatokról az összetett mondatoknál írok.

5.3.3 Vonzattranszformáció

A szavakhoz tartozó vonzatok megadása szótárazási munka. Emiatt jó, ha minél kevesebb adattal sok szónak tudjuk megadni a vonzatát. Az előző fejezetben két képzőnek meg is adtuk általánosan. A leggyakoribb képzés az igenevek előállítása. Ha van általános algoritmus, hogy egy igéből képzett szó vonzatait hogyan kapjuk meg az eredeti szóból, akkor nyert ügyünk van. Nos, ez nem is olyan nehéz. Néhány példa:

Folyamatos melléknévi igenév: Az alanyon kívül az ige minden vonzata megmarad, és az ige alanya lesz az igenév által jelzett szó:

Péter hosszú távot fut le mindennap.
A hosszú távot futó Péter…

Az *–andó/endő* melléknévi igenév: Csak tárgyas igékre lehet használni. Az alanyon és tárgyon kívül az ige minden vonzata megmarad, és az ige tárgya lesz az igenév által jelzett szó:

Péter által lefutandó táv…

Az *–ható/hető/hatalan/hetelen* melléknévi igenév: Csak tárgyas igékre lehet használni. Az alanyon és tárgyon kívül az ige minden vonzata megmarad, és az ige tárgya lesz az igenév által jelzett szó:

Péter által lefutható táv… Péter által lefuthatatlan táv…

Befejezett melléknévi igenév: Az alanyon és tárgyon kívül az ige minden vonzata megmarad, ha az igenév által jelzett szó az ige tárgya, illetve az ige alanya *által* névutós szerkezetté válik, mint az igenév nem kötelező vonzata. A nem tárgyas igéknél a jelzett szó általában az ige alanya, tárgyas igéknél a tárgya szokott lenni, de lehet az alany is. Alternatív vonzatkeret, ha a jelzett szó az ige alanya:

A Péter által lefutott hosszú táv …
A vert török sereg…
A sokat játszott színdarab…

A hosszú távot lefutott Péter…
A sokat látott bölcs…

Ezek a megállapítások igazak akkor is, ha a befejezett melléknévi igenév tagadó alakját használjuk, bár az *által* vonzatot nem használjuk:

A Péter által le nem futott hosszú táv …
A veretlen focicsapat …
A lejátszatlan sakkmeccs …
A kialvatlan felnőtt …
A hosszú távot futott Péter …

Műveltetés: Az alanyon kívül az ige minden vonzata megmarad, illetve az eredeti ige alanya *-val* ragos szerkezetté válik mint az igenév nem kötelező vonzata:

A szabóval új ruhát szabattam …

Az igék egy csoportja, főleg a mozgást jelentő igék attól kapnak vonzatot, hogy igekötővel rendelkeznek:

Keresztülnéz az ellenségen.
Átmegy a hídon.
Utánanéz az ügynek.
Kiűzi az országból…
Begyömöszöli a zsebébe…
Bement a szobába…

Ezekben a példákban az igekötő határozza meg az estlegesen kötelező vonzatot. Néha az igének nincs is igazi jelentése, csak a ragokat hordja, és az igekötő a jelentést hordozó része a szónak, esetleg az igazi jelentést nem az ige, hanem az igekötő viseli. Legalábbis a vonzatkeretet az határozza meg:

Bekönyörgi magát a rendezvényre…

Sajnos ez általánosan nem használható vonzatmeghatározó algoritmus. Gyakran az igekötő lényeges jelentésváltozást hoz, ezzel együtt vonzatkerete is lényegesen eltér az eredeti igéétől. Ilyenkor csak szótárba vétellel lehet kezelni az esetet. A következő példákban az *ad* ige jelentéséhez vajmi kevés köze van az igekötős igének.

Megadja magát	*Kiadja a lelkét*	*Átadja magát a zenének*
Feladja magát	*Beadja a kulcsot*	*Eladja az utolsó adogatást*
Ráadja a ruhát	*Leadja az órát*	*Összeadja a költségeket*

Feladat 1: Nézze át a képzőket, és állapítsa meg, melyeknek lehetnek saját vonzatkeretük, és melyek hogyan változtatják a vonzatokat.
Feladat 2: Nézze át az igekötőket, és állapítsa meg, hogyan módosul a mozgást jelentő igék vonzatkerete.

5.4 Szabad határozók

A mondatban sok olyan mondatrész is szerepelhet, melyek látszólag sehova nem köthetők, vagy bármely, főleg igei tulajdonságú részekhez csatolhatók. Ezek a szabad határozók. A leggyakoribb szabad határozók idő- és helyhatározók. Hogy formálisan hogyan különböztessük meg az idő és a helyhatározókat egymástól, nem könnyű feladat. A legtöbb rag, névutó mindkettőt jelölheti. Ha viszont némi szószemantika is van a háttérben, akkor az időhatározók nagy biztonsággal kiválaszthatók.

A Dunán evezek minden szombaton.
Szilveszter éjszakáján otthon ünnepeltünk.

de: *Egy órán belül megérkezem – Egy órán belül sok érzékeny alkatrész van.*

Vonzatnak azért nem nevezhetném, mert ha nem is bárhol, de elég sok helyen egészítheti ki a mondatot. Az, hogy minek az idejét helyét jelöli, bizony nehéz megállapítani egy mondaton belül.

Az idő- és helyhatározónak is alkategóriái a honnan, hova, hol, milyen távon, illetve a mikortól, meddig, mennyi ideig…

Társ- és eszközhatározók – magyarban szemantikai ismeretek nélkül e kettőt nehéz elkülöníteni:

Péterrel ette meg a kenyeret – Szalonnával ette meg a kenyeret – Kanállal ette a levest.
Péter nélkül ette meg a kenyeret – Szalonna nélkül ette meg a kenyeret – Kanál nélkül ette a levest.
Egyedül ette meg a kenyeret – Magában ette meg a kenyeret.

A mennyiséghatározók – ezek formálisan tárgyesetű szavak:

Ha hármat alszol, itt a Karácsony.

Sokat mérgelődött az eseten.
Az alma már piros egy kicsit

Ha az ige tárgyatlan, vagy névszói mondatról van szó, akkor a tárgyesetű kifejezés valószínűleg mennyiséget határoz meg. Más esetben kérdéses:

Várt egy kicsit – és egy ilyen nagy jött helyette. (Prószéky Gábor példája)

De nem ritka a gyakoriséghatározó, mely nem mindig az időhatározók alkategóriája, de jól felismerhetők.

Hetente kaptam levelet.
Nem megmondtam százszor?
Ritkán utazom külföldre.
Bejáratonként egy-egy őrt állított.
Péter gyakran megy egyedül színházba.

Az utóbbi mondatban a *színházba* a mozgást jelentő ige vonzatának tekinthető, míg az *egyedül* társhatározó és a *gyakran* gyakoriséghatározó, mindkettő tipikus szabad határozó, a *színházba* a *megy* ige nem kötelező vonzata.

Gyakoriak még a módhatározók is – pontosabban a cél-, ok-, állapot- és módhatározók. Hát ezek azok, amelyeket a legnehezebb felismerni. Néhány névutó, névmás egyértelműen ide sorolja a névszói kifejezést, más esetekben a nyelvészek olyan képtelen kategóriát használnak, mint a „képes helyhatározó”, mert formálisan helyhatározói raggal, névutóval rendelkeznek.
Egyéb kategóriát is felállíthatunk, mint pl. a szerint.

A meteorológiai központ jelentése szerint eső nem várható.
Állítólag már elfogták a drogkereskedőt.

vagy a helyett

Péter helyett Pál jött el.
Villa helyett kanalat vett elő.
Mozi helyett színházba mentünk..

mely viszont sohasem globális, az egész mondatra vonatkozó határozó, hanem általában valamely mondatrészre vonatkozik. Az is lehetséges, hogy a helyett–et a fókuszban lévő mondatrészhez kell társítani (lásd: **téma réma, megjegyzés**).
A névszói kifejezésekben igenevek kiegészítőjeként is megjelenhet, akár az ige egyéb vonzatai.

A gyakran egyedül az utcán csellengő gyerekek veszélynek vannak kitéve.

5.5 Egyszerű mondatok típusai

A nem összetett mondatok angolban mindig tartalmaznak egy központi igét. Nem így a magyarban és például a szláv nyelvekben. Az igét tartalmazó mondatoknak is két fő típusa van. Az egyik a létezést kifejező, a másik a szokásos cselekvést, történést kifejező főigés mondatok. Formailag az ige(i kifejezés) a mondat feje, központja.

Almát eszem.
Péterék tegnap csak almát vacsoráztak.
Nem ehetik meg az almát mosatlanul.
Lassan mindent rosszat elfeledtet az idő.

A létezést kifejező ige köré épülő mondatoknál a kérdéses ige a *van* (*lesz*, *nincs*). Ha így tekintjük, akkor ennek az igének egy kötelező vonzata van, az alany.

Tizenhárom fodor van a szoknyámon.
Egy, csak egy legény van talpon a vidéken.
Minden rendben van.
A héten nem volt ötös a lottón.
Hol voltál báránykám?

Érdemes külön tárgyalni a birtokviszonyt kifejezőket.

Nincsen apám, se anyám, se istenem, se hazám, se bölcsőm, se szemfedőm, se csókom, se szeretőm.
Nekem senkim sincsen.
Nagy bánata van a cinegemadárnak.

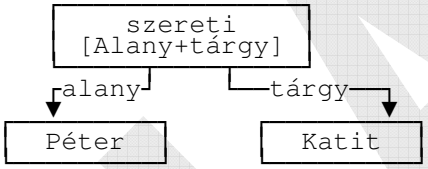
A nem igei mondatok érdekessége, hogy jelen idő kijelentő mód harmadik személy estén nem használunk semmilyen igét, hanem két alanyesetű névszói szerkezet közti viszony adja a mondat lényegét. Ha csak egy alanyesetű névszói kifejezésből áll a mondat, ezt csak hiányos mondat lehet. Hát nézzük részletesen:

5.5.1 Igei mondatok

Az igei mondatok, teszik ki mondataink nagyobbik részét. Tulajdonképpen változatos típusúak. A típust alapvetően a főige – mint a mondat feje – határozza meg. A vonzatkeretnek megfelelő vonzatok szerepelhetnek benne, kiegészítve szabad határozókkal. A főige és az esetleges segédigék határozzák meg a mondat idejét és módját. Ettől teljesen függetlenül tölthetjük ki a vonzatkeret egyes vonzatait határozószavakkal, névszói kifejezésekkel. A vonzatok mikéntjétől független azok sorrendje a mondatban. Szabad határozók is szerepelhetnek a mondat különböző helyein. A vonzat egyes elemeinek is lehetnek vonzatai, és ha a vonzat nem kötött sorrendet követel, akkor ezek „el is vándorolhatnak" attól a szótól, amihez tartozik. A vonzat vonzatának is lehet elvileg vonzata, de ez a gyakorlatban ritkán fordul elő. Miután az ige vonzatainak sorrendje általában szabad, és a vonzat vonzata sem kell kötelezően a vonzott szó mellett állnia, a szósorrend leírására alkalmas környezetfüggetlen nyelvtan nehezen használható.

A klasszikus három szóból álló mondat mind a hat szósorrendjében a függőségi összefüggés, melyet címkézett gráfban ábrázolhatunk, azonos:

Péter szereti Katit. *Péter Katit szereti.* *Katit Péter szereti.*
Katit szereti Péter. *Szereti Katit Péter.* *Szereti Péter Katit.*

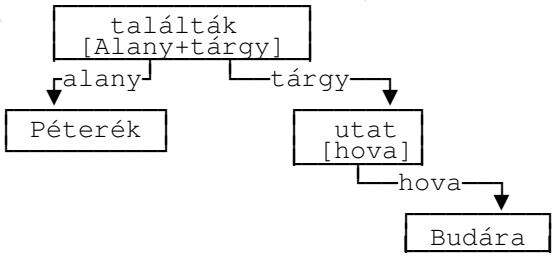


A gráf arra inspirál, hogy ezt egyszerű környezetfüggetlen nyelvtanból is megkaphatjuk, hiszen ott is hasonló levezetési fát kapunk. A környezetfüggetlen nyelv levezetési fájában az egy csomópontból kiinduló éleket nem címkézzük, hanem csak a csomópontokról mondjuk meg, hogy milyen nyelvtani, esetlegesen terminális szimbólumnak felel meg, és azt, hogy az onnan kiinduló élek adott sorrendben melyik nyelvtani szabálynak felelnek meg. Tehát az élek minősítését az élek sorrendje határozza meg. Ha ez megoldható – nyelvészi szóhasználattal – a szintaxis konfigurálható. Angolban és sok más kötött szórendű nyelvnél így van. Ha a fenti példát is hasonlóan szeretnénk megoldani, akkor a következő nyelvtantöredéket kellene használni:

<Igei mondat>→<igei kifejezés> <alany> <tárgy> | <igei kifejezés> <tárgy> <alany> |
<alany> <tárgy> <igei kifejezés> | <alany> <igei kifejezés> <tárgy> |
<tárgy> <igei kifejezés> <alany> | <tárgy> <alany> <igei kifejezés>

Ennek megfelelően a hat mondatnak hat, egymástól különböző fáját kapjuk meg elemzéskor, melynek persze lehet egységes címkézése. Ezzel megoldhatjuk a gondot, de ha a másodlagos elváló vonzatot is szeretnénk hasonlóan kezelni, akkor már bajban vagyunk:

Budára nem találták Péterék az utat



A fenti mondatnak négy mozgatható része van, ezért azonos függőségi relációval 4!, azaz 24 szabályos sorrendje létezik. Nem is a nagy szám itt a fő gond, hanem az, hogy egy alsóbb szint, a tárgy egy kiegészítője, a hova része nem a tárgy mellé van beágyazva. Emiatt ez a mondat nem elemezhető úgy környezetfüggetlen nyelvtannal, hogy a levezetés a függőségi kapcsolatot tükrözze.

5.5.2 Egzisztenciális mondatok

A létezést kifejező mondatok, mint korábban kifejtettem, kötelezően egy létigét és egy alanyesetű kifejezést tartalmaz, esetleg szabad határozókkal bővítve a mondatot. Ha úgy fogjuk fel, hogy a létigének egyargumentumú kötelező vonzata van, akkor algoritmikusan visszavezettük a feladatot az egyéb igés

mondatokhoz. Azt, hogy az idő- (mikor) és (hol) helyhatározót szabad határozónak, vagy nem kötelező vonzatnak tekintjük, megfontolható az alkalmazás szempontjából. Magam sem tudok itt dönteni.

Nincsen apám, sem anyám
Holnap lesz vasárnap
Van benne valami

Különleges eset, amikor a létige mégsem kötelező – egy mutató névmás mellett áll az alany, és a létigét elhagyhatjuk:

Itt a nyilam, hova lőjem.
Ott egy legény a folyó partján

A szenvedő szerkezetű mondatokban, ha az határozói igeneves, felmerül a kérdés, hogy hova soroljam. Lehet szerkezetileg egzisztenciális mondatként is elemezni, melyben egy szabad módhatározót találunk, de praktikusabb az igenévvel együtt a létigét „elváló" segédigés szerkezetnek tekinteni, figyelembe véve a vonzattranszformációkat (lásd, **5.5.5**):

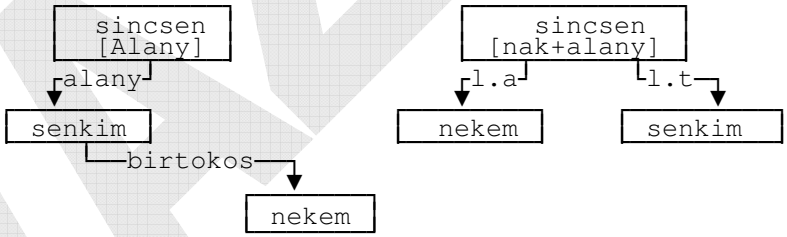
Zöldre van a rácsos kapu festve.
Nem voltunk felkészülve az árvízre.

5.5.3 Birtokviszonymondatok

A szenvedő szerkezetű mondatokhoz hasonlóan érdemes külön kezelni a birtoklást kifejező mondatot. Ha megfigyeljük, minden esetben egy esetlegesen elváló birtokos szerkezet létét lehet felfedezni egy egzisztenciálismondatban.

Mázlija van
Nekem senkim sincsen

Azért érdemes egy külön kategóriába sorolni, mert a legtöbb indoeurópai nyelvben erre külön ige van, és a mondatszerkezet emiatt lényegesen más, mint a magyarban. Például az angol, francia birtokos lesz a mondat alanya, míg a magyarban mindig egy *-nak* ragos névszói kifejezés adja a birokost. Furcsa módon érezzük is a birtokos „alanyságát", és sok esetben – főként a mondatokon áthúzódó összefüggésekben – úgy viselkedik, mint más mondatokban az alany szokott. Emiatt mondjuk is, hogy ezekben a mondatokban a logikai alany a birtokos. A két függőségi gráf eltér, és logikusabbnak tűnik a második, ahol a birtokos és birtok egy szinten van:



5.5.4 Névszói állítmányos mondatok

A nem igei mondatok érdekessége, hogy jelen idő kijelentő mód harmadik személy estén nem használunk semmilyen igét, hanem két alanyesetű névszói szerkezet közti viszony adja a mondat lényegét. Ha csak egy alanyesetű névszói kifejezést találunk létige nélkül, ebben az esetben hiányos mondatot eredményez. A két névszói kifejezés valamilyen viszonyban van. Ez talán már a szemantikába hajlik, de mivel formai ismérvek alapján is lehet különbséget tenni, ezért érdemes a szintaxist is felhasználni.

Három jellemző dolgot lehet kifejezni névszói mondattal.

1. Valami valamilyen — *Új a csizmám*
2. Valami valamilyen kategóriába (halmazba) tartozik. — *A kutya emlősállat*
3. Két dolog ekvivalens — *A legszebb te vagy*

Az első kettőben meg tudjuk különböztetni, melyik alanyesetű kifejezés az alany, melyik az állítmány. A harmadikban erre nincs mód. Mindhárom esetben a létige nem jelenik meg, ha kijelentő mód harmadik személyű az alany más estekben is csak azért, hogy kifejezzük az állítás módját, esetleg múlt idejét, illetve nem harmadik személyű alany esetén az ige szokta hordozni az alany számának, személyének ragját.

Új volt a csizmám. A kutya emlősállat lehet A legszebb te vagy

Az első esetben alanyesetű határozatlan jelzői szerkezet van egy alanyesetű határozott névszói kifejezés mellett, a másodikban egy határozatlan névszói kifejezés van egy határozott mellett, a harmadikban pedig

két alanyesetű határozott kifejezést találunk. Most egy új nyelvi terminust vezettem be: határozott, és határozatlan. Ez két helyen szerepel a mondattanban. Az egyik itt, a névszói mondatoknál, másik pedig a tárgy és az igeragozás egyeztetésénél. Sajnos a két esetben eltér a jelentése. Nos, itt azt jelenti, hogy lehet-e alanya egy névszói mondatnak. Az a tapasztalat, hogy ebből a szempontból akkor határozott, ha az **5.1.1.** fejezetben említett névszói szerkezetben a szelekció-t, vagy annál balrább lévő kategóriát tartalmaz a kifejezés. Például ha határozott névelője, mennyisége vagy birtokosa is ki van fejezve a névszói kifejezésben.

Sok lúd sánta volt.
Kutyája is szelíd fajta.
Egy legény sem lehet elég bátor.
Nem volt a Szása egy moszkvai nagy dáma.

De nem lehet jó.

** Lúd sánta volt.*
** Kutya is szelíd fajta*
** Legény sem lehet elég bátor.*

Ha mindkettő névszói kifejezés határozott, akkor formailag nehéz megállapítani, melyik az alany, melyik az állítmány:

A katona a gyilkos
Italuk a száraz martini.

Emiatt inkább az a logikus, hogy két dolog ekvivalenciáját fejezi ki. Hogy az első mondatban például a *katonáról* állítjuk, hogy *gyilkos*, vagy a *gyilkos*ról, hogy *katona*, az magyarul nem derül ki. Ha angolra, franciára, németre fordítanánk, mégis kényszerülünk megkülönböztetni az alanyt az állítmánytól. Ilyenkor egy alkalmas eljárás, hogyha azt nézzük, a névszói kifejezésben melyik ért el magasabb kategóriát az **5.1.1.** fejezetnek megfelelően, ebben az értelemben melyik van jobban specifikálva. Más módszer is van, de azt az **5.8.** fejezetben tárgyaljuk.

Logikailag az első kettő mondatkategóriát akár össze is vonhatnánk. Az egyszerűsített logikában egy tulajdonság ekvivalens az általa definiált karakterisztikus halmazzal.

5.5.5 Igeneves (szenvedő) mondatok – Hova is soroljam 1.

A birtokviszonyt kifejező mondatok külön kategóriába való vételéhez hasonló a helyzet a szenvedő szerkezetes mondatoknál. Szenvedő szerkezetű mondatok két fajtája van. Az egyik, amikor tárgyas ige befejezett melléknévi igeneve egy valami valamilyen jellegű mondat állítmánya, a másik, amikor a tárgyas ige határozói igeneve csatlakozik egy formálisan egzisztenciális mondatban a létigéhez. Ezeket azért érdemes külön kategóriába venni, mert ha ilyen jellegűek a mondatok, a mondattranszformációkat betartva kötelezők maradnak a vonzatok, szemben azzal az esettel, amikor csupán jelző, illetve határozó válik az igéből.

Zöldre van a rácsos kapu festve. *A rácsos kapu frissen festett.*
A hal enyhén meg van sózva *A hal enyhén sózott.*
Ezzel a feladat meg lett oldva. *Ez a feladat már megoldott.*
Nincs kizárva, hogy… *Kizárt, hogy…*

Vegyük észre, hogy abban az esetben, ha az igekötő szerepe csak az, hogy az ige befejezettségét jelentse, akkor a befejezett melléknévi igenévnél nem szerepelhet. Az igenév önmagában jelzi már a befejezettséget. *A katona lelőtt* nem jelentheti azt, hogy a *katonát lelőtték*, csak azt, hogy *a katona lőtt le valakit*.

Ha a névszói mondatban az állítmány egy folyamatos melléknévi kifejezés akkor is felmerül a kérdés, külön mondattípusnak tekintsük, és betartva a vonzattranszformáció szabályait igei mondatként elemezzünk, vagy sem.

5.5.6 Infinitívuszos mondatok – Hova is soroljam 2

A segédigés szerkezetekhez hasonlóan viselkedik néhány melléknév is, Ezek között melléknevek között is megvan a kétféle, az egyikben az infinitívusz kaphat személyragot, a másikban nem.

Péter képes megoldani a feladatot. *Képes vagyok megoldani a feladatot,*
Tilos az ablakon kihajolni. *Tilos az ablakon kihajolnod.*
Szabad péntek, szabad szombat, szabad szappanozni,

Itt sokan alanynak vagy tárgynak érzik az igenevet, és nem segédigés szerkezetnek a mondatot. A melléknév ebben az értelemben az állítmány. Mivel az én beosztásom szerint az infinitívusz nem névszó, sok szempontból nem érdemes annak tekinteni. Itt is inkább segédigés mondatsémába sorolnám ezt a típust, hogy a főige, mely most infinitívuszi alakot vesz fel, hogy az elemzés szerint a mondat központjába kerülhessen.
A fenti dilemma olyan mondatoknál élesebben jelentkezik, ahol egzisztenciális mondathoz hasonló szerkezetben lenne az alany a hagyományos elemzés alapján az infinitívuszos szó:

Nincs mit tenni,
Van hova mennie
Péternek nem volt kivel beszélnie.

Ezekben az esetekben látjuk, hogy a létigét az infinitívuszossal együtt, mint főigével egy igei csoportot alkotva logikusabban elemezhető szerkezetet kapunk, mintha egyszerűen egzisztenciális mondatnak tekintenénk.

5.5.7 Egyeztetések a mondatrészek között

A magyarban a mondatrészek bizonyos tulajdonságai más részektől függenek. Itt most nem a szófaji beosztásra gondolok, hanem például arra, hogy az alanynak számban, személyben meg kell egyeznie az igerag számával, személyével. Ez az alapvető egyeztetés a legtöbb nyelvben megvan. E nélkül agrammatikus a mondat – bár érthető lehet – és megértéssel viseljük, ha egy külföldi beszédben vét ez ellen a szabály ellen. A külföldiek többsége a tárgyas ragozást is elvéti. Ilyen tulajdonság kevés nyelvben van. Viszont van olyan finnugor családhoz tartozó nyelv, ahol az igerag a tárgy személyét, esetleg számát is magába foglalja. Ezeknél pontosabb egyeztetés van az igerag és a tárgy között. (lásd **4.10.4.**) Ha az alany egyes szám első személyű, a tárgy második személyű, akkor ez az igeragban explicit szerepel.

Nem látt<u>alak</u> téged a színházban tegnap este.

Érdekes eset, ha halmozott mondatrész az alany vagy a tárgy. Formálisan a következő a szabály. A halmozott alany személye az egyes elemek személyének minimuma. Ha az egyes elemek harmadik személyűek, és egyes és többes számú is szerepel a felsorolásban, akkor az ige száma lehet egyes és többes számú. Ha első és más személyű is szerepel az alanyi részben, akkor az igerag többes számot kap.

Én és Pista gyakran találkozunk.
Te és Juli nem találkoztok.
Juli és Pista soha nem találkozott. Juli és Pista soha nem találkoztak.
Pista és a diákok gyakran találkoztak.

Ha a tárgy harmadik személyű, akkor más igeragot kap az ige aszerint, hogy határozott névszói kifejezés a tárgy, vagy nem. Itt a határozottság definíciója eltér attól, mint amit a névszói mondatoknál használunk. A mennyiség még nem teszi határozottá a kifejezést az ige szempontjából.

Itt határozott a *három*: *Három sok nekem.*
Itt határozatlan: *Hármat veszek.*

Más, egyeztetésként is felfogható tulajdonság, hogy mennyiséget tartalmazó névszói kifejezést nem lehet többes számban használni. Ez az angolszász és latin nyelvekkel ellentétes, mert ott, ha nem egy a mennyiség és nem megszámlálhatatlan, akkor számnév után többes számot kell használni. A szláv nyelvekben viszont összetettebb szabályok vannak – egyes, illetve többes számú birtokos esetet használnak. Eltérő szabályok vannak a páros jelentésű főnevekre – *szem, fül, kéz, cipő*, sok nyelvben a *nadrág,* a *szemüveg* is ilyen. A magyartól eltérően többes számot használnak az indoeurópai nyelvekben, és ha csak egyről, mondjuk, pl. a *bal kesztyűről* van szó, akkor használnak egyes számot. Magyarban ellenben a párt általában csak egyes számban használjuk, és ha csak egyikről van szó, akkor helyesebb a

fél fülére süket, *nem találja a fél zokniját*, *fél lábon áll…*

kifejezéseket használni, bár helyesek az

egyik fülére süket, *nem találja az egyik zokniját*, *egy lábon áll…*

mondatok is.

Egyeztetés a birokviszonynál a birtokos birtokragjának és a birtoknak megfeleltetése is. Itt egy érdekes jelenség van. Ha a birtokos az *ők* személyes névmás lenne, akkor csak egyes számban használjuk, akkor is, ha a birtokrag többes szám harmadik személyt jelez.

Ez nem az ő kutyájuk.

Egyesek a vonzatot (a vonzott rag, illetve névutó kapcsolatát) is egyeztetésnek nevezik. Az adott formalizmus dönti el, minek nevezzem. Esetleg a névutó és a névutóhoz tartozó névszói eset viszonyát magam is a vonzatként kezelném. Összefoglalva, a következő főbb egyeztetések vannak a magyarban:

1. Ige(rag), és az ige alanyának száma személye.
2. Ige(rag), és az ige tárgyának határozottsága, esetleg személye.
3. Egyes birtokragos kulcsszavas vonzatok, (pl. visszaható névmás) és az alany (esetleg más vonzat) száma és személye.
4. Birtokos és birtok birtokos jelének száma, személye.
5. Mennyiség és egyes szám használata.

És amiket technikai szempontból nem egyeztetésnek, hanem inkább vonzatnak neveznék:

6. Névutó vonzata és a hozzátartozó névszói kifejezés esete.
7. Mellékmondat módja, amelyről az alárendelt összetételnél írok
8. És egyéb vonzatok…

5.5.8 Címmondatok

Helyes mondatok, pontosabban mondatszerű kifejezések, melyek írásokban megjelennek, a címmondatok. Mint neve is mutatja, többnyire címekben fordulnak elő olyan ige nélküli mondatok, melyek sehova se férnek a fenti leírás alapján. Vagy egy, vagy két nem feltétlen határozott alanyesetű névszói kifejezést tartalmaznak, mindenféle ige nélkül:

Gyilkos az aluljáróban
Ábel Amerikában
Szoba kiadó
Súlyos lakhatási problémák

Ezek szinte kizárólag címekben, feliratokban, reklámszövegekben, szlogenekben fordulnak elő, de mivel ilyeneket is kell elemezni, fordítani, fel kell rájuk készülni.

5.5.9 Mondatszók, indulatszavak

A szófaji beosztásban nem szerepel, de létezik olyan szófaj, amelyhez tartozó szavak önmagában mondatnak tekinthetők. Ezek szótanilag nem okoznak gondot – ragozhatatlanok. Mondattanilag pedig teljes értékű mondatnak tekinthetők. Ilyenek:

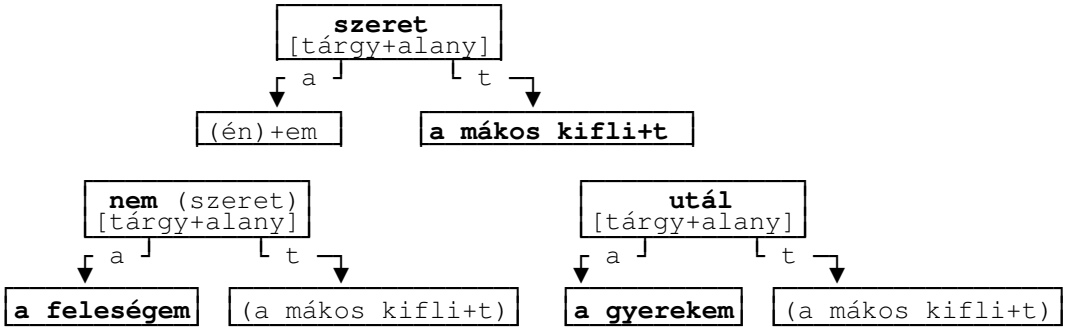
indulatszó: *óh, jaj, pfuj, hű*
mondatszó: *persze, sajnos, sajna, hopp*

5.5.10 Hiányos mondatok

A magyar nyelv különös mértékben elviseli, ha egy mondatból akár lényeges rész is hiányzik. Ha önállóan elemzünk egy mondatot, akkor egyes – akár kötelező – vonzatokat is elhagyhatunk. Sőt, a mondat feje, az ige is hiányozhat:

Szeretem a mákos kiflit. A feleségem nem. A gyerekem utálja.

Az első egy teljes mondat, a másodikból az ige hiányzik, a harmadik viszont az értelmezésem szerint nem hiányos, mert az ige tárgyas ragozása magában foglalja valamilyen mértékben a tárgyat, de más felfogásban hiányos. Hasonlóan, az első mondatban az alany be van olvasztva az igébe.



Angolul mindkét esetben különálló névmást használunk. A második mondat angolra fordításánál egy „igemás”-t kell használni, a *do* segédigét. Mert az angol nyelv, talán a kötött szórend miatt, nem viseli el a hiányt.

I like the puppy seeded croissant. My wife does not. My child hates it.

5.6 Összetett mondatok

Nem definiálom, mi az összetett mondat. Az anyanyelvű beszélő többé-kevésbé tudja. De honnan tudja a gép? Az biztos, hogy az összetett mondat több egyszerű mondatból áll. A szintaxis szempontjából nem elég, hogy a tagmondatok összege, mert az összefüggés nem egyszerűen a részek együttese. A tagmondatok közötti viszony lényeges.

A magyarban, mint más nyelvben is, formális jele van az összetételnek. Az biztos, hogy a tagmondatok közé vesszőt (esetleg pontosvesszőt, kettőspontot) kell tenni, még akkor is, ha kötőszót használunk. Mondatainkban a vessző viszont ugyanúgy lehet halmozott mondatrészek között, mint tagmondatok között. Ez okozhat gondot. Egyes esetekben viszont tagmondatokat elválasztó, koordináló kötőszavak szerepelnek, amelyek biztosíthatják, hogy csak összetett mondatról van szó. Ez azért fontos, mert e mankó nélkül az elemzőnek fáradságos holt utakat kell bejárnia, hogy kipróbálja, melyik alternatíva a helyes, melyik vesszőt értelmezze a gép részmondat határának, és melyek jelentenek csak halmozott mondatrészek közötti elválasztót. Ha sikerül a tagmondatokra való szegmentálás, akkor az elemzőnek sokkal könnyebb dolga van – marad így is mit megoldani.

Ha egy mondat nagybetűvel kezdődik; ponttal, felkiáltó- jellel vagy kérdőjellel végződik, és egy vessző van benne, továbbá a vessző két oldalán egy-egy ragozott ige áll a saját vonzataival; akkor többé-kevésbé világos, hogy itt összetett mondatról van szó, amelynek két tagmondatát a vessző választja el. Nem minden mondat ilyen, sajnos. Pl.:

Egész nap az ágyban evett, ivott, heverészett.

Ennél a mondatnál valószínűleg helyesebb, ha úgy tekintjük, hogy összetett (halmozott) állítmánya van, és az ehhez kapcsolódó hely és idő a benne szereplő összes tevékenységre vonatkozik. Nem biztos, hogy ugyanez a helyzet annál a mondatnál is, hogy:

Egész nap evett, ivott, heverészett az ágyban.

Az *egész nap* mindháromra vonatkozik, *az ágyban* pedig lehet, hogy mindháromra, de lehet, hogy csak a *heverészés*-re. Megint más a helyzet, ha azt mondjuk, hogy:

Egész nap evett, ivott, az ágyban heverészett.

Az *ágyban* itt egyértelműen csak a *heverészés*-re vonatkozik.

Egy lehetséges megoldásnak látszik, ha a mondatot a vesszőknél úgy vágjuk szét, hogy minden egyes részbe egy ragozott ige kerüljön, és az így kapott mondatokat elemezzük. A kapott mondatok természetesen valószínűleg roppant hiányosak lesznek. Nem látszik túl egyszerűnek annak eldöntése, hogy a hiányok közül mit lehet pótolni az előző tagmondatokból öröklődéssel. A megoldásban nyilván fontos szerepe lesz a topiknak és a fókusznak, és a megoldás nem lehet tökéletes, mert a hangsúlyt írott szövegnél nem tudjuk figyelembe venni. Más oldalról is felmerül a kérdés, honnan ismerjük meg a mondat azon részét, amit önálló tagmondatnak kell tekinteni, milyen mondatrészeknek kell abban szerepelniük.

Máskor is gyere hozzám, de este.
Ma is azzal láttam Pistát, akivel tegnap.
Nem csak én mondom, hanem mindenki.
Pestnek jövője van, Budának múltja.
Már nagyocska gyermek voltam, olyan ötödfél esztendős.
Akkor szép az erdő, mikor zöld.
Kicsi a rózsám, de csinos.
Addig hajlítsd a fát, amíg fiatal.

Noha az összetett mondatok 80-90%-a viszonylag normális mondat, van állítmánya, ugyanakkor az sem ritkaság, ha az egyik tagmondat (rendszerint a második) olyan csonka, hogy önmagában elemezhetetlen. Az igei állítmányú mondat rendszerint azért, mert nincs ott az ige, nincs mihez kötni a vonzatokat; a névszói pedig rendszerint azért, mert egyetlen alanyesetű névszói kifejezés van benne ige nélkül. Az esetek igen nagy részében ezen megpróbálhatunk segíteni. Ha az első mondat igei állítmányú, és a második olyan

hiányos, hogy elemezhetetlen, akkor megkísérelhetjük úgy elemezni, mintha az első mondat igéje állna benne. Ha az első mondat névszói, akkor feltehetjük, hogy a másodikból az első tagmondat alanya hiányzik.

Persze ez az eljárás nem mindig válik be. Például az utolsó példamondatra sem alkalmazható. Az is elképzelhető, hogy félreértés lesz belőle. Bár ilyenkor rendszerint a kiinduló mondat is kétértelmű volt, legalábbis nyelvtanilag.

Van egy mondatfajta, az összehasonlítás, ami majdnem mindig hiányos. Legalábbis "A magyar helyesírás szabályai" és "A mai magyar nyelv" egyes fejezetei explicite azt mondják, hogy az összehasonlító mondatokban a *mint* kötőszó utáni rész egy önálló mondat. Mivel az a tapasztalat, hogy az esetek 95%-ában egyetlen névszói kifejezés áll a *mint* után, további 4%-ban pedig vonatkozó névmással bevezetett almondat, ne így fogjuk fel a dolgot, mivel nehézkes hiányos mondatokat elemezni.

Két fő típusa van az összetételnek: alá- és mellérendelő. Az alárendelő összetételben az egyik tagmondat kiegészíti a főmondatot, gyakran egy mondatrészét fejti ki részletesen, netán a főmondat igéjének egy vonzata az egész mellékmondat. A mellérendelésnél a két mondat állításának a viszonyát a kötőszó határozza meg, de logikailag a két mondat egyenrangú.

5.6.1 Alárendelő összetétel

Az alárendelt összetételnek két fajtája van. Formálisan a főmondat egy részét fejezi ki vagy pontosítja a mellékmondat. Az első esetben az almondat egy mondatrésze felel meg a főmondat egy mondatrészének, a másikban az egész almondat állítása a főmondat egy mondatrészének a kifejtése.
Az első esetben a főmondatban egyik mondatrész csupán egy mutató névmás, a mellékmondatban pedig egy vonatkozó névmás szerepel, és ezek ugyanazt az objektumot jelentik, illetve hasonló szerepet töltenek be mindkét mondatban:

Azt ette, amit kapott.
Balsors akit régen tép, hozz reá víg esztendőt
Olyan ételt nem eszem, amelyik húst tartalmaz.
Úgy aludt, ahogy még sohasem korábban.

Ez utóbbi mondatban az *úgy* szót mutató névmásnak tekintem – távolra mutató módot jelölő névmás, az *ahogy* pedig módot jelölő vonatkozó névmásnak felel meg. Hasonlóan az *ott – ahol* pár helyhatározói mutató, illetve vonatkozó névmás.
Hol a mutató, hol a vonatkozó névmás marad el. A mutató névmás akkor szokott elmaradni, ha a két mondatban hasonló funkcióban, vonzatban szerepel az azonos, összekapcsolható mondatrész. A vonatkozó névmásoknak gyakran az *a-* nélküli alakját használjuk:

(A)Mit rákentek a századok, lemossuk (azt) a gyalázatot.
(Azt) ette, amit kapott.
(Úgy) aludt, ahogy még sohasem korábban.

A *mint*-es mellékmondatok egy része is ilyennek tűnik. A legtöbb esetben az *olyan*, *annyival* és hasonló mutató névmások hiányoznak a főmondatból, és a mellékmondat erősen hiányos, illetve a mellékmondatot mindig megtoldhatjuk egy vonatkozó névmással is, ha hiányzik. Ez mutat arra, hogy a *mint* nem biztos, hogy ide tartozik, hanem egy hiányos második típusú mellékmondatot vezet be, amely egy beágyazott almondatot is tartalmaz. Ennek ellenére az alábbi összetételekben a *mint* nehezen értelmezhető a második típusú alárendeltség kötőszavaként.

Olyan csökönyös, mint (amilyen csökönyös) egy szamár
Sokkal többet dolgozik (annál), mint amennyit elvárnak tőle.
Nagyobbat ugrott (annál), mint (amilyen nagyot) Péter (ugrott).
Nagyobbat ugrott (annál), mint (amilyen nagy) Péter.

A mutató és vonatkozó névmást tágan kell értelmezni. Vannak általános mutató névmások, melyek formailag nem származtathatók az *az,* illetve *ami* névmásból: Ilyenek a helyhatározó *ott - ahol*, *oda - ahova*, *onnan - ahonnan*, vagy a módhatározó *úgy – ahogy* párok.

A másik típusnál, amikor az egész almondat a főmondat egy részének kifejtése, akkor is lehet formálisan a főmondatban egy mutató névmás, de a mellékmondatban ilyenkor nem vonatkozó névmás szerepel, hanem a *ha*, *mert, minthogy, mint hogysem* vagy a *hogy* kötőszó vezeti be. A főmondatban ilyenkor is gyakran használunk mutató névmást. A *ha* esetén az időhatározó *akkor*-t, de a *hogy* esetén sok más is lehet. A *ha*-s mellékmondatot különös időhatározói mellékmondatnak nevezik egyes nyelvészek, de inkább a főmondat állításának előfeltételét jelenti, ami nem feltétlen jelent előidejűséget:

Ha a piros kabátot látod a fogason, akkor Kati már megérkezett.

Tehát a *ha* és a *mert* feltétel-, illetve ok-, a *hogy* pedig célhatározói mellékmondatot vezet be, vagy csak egyszerű, a korábbi kategóriába tartozik, ahol a főmondat mutató névmásának kifejtése a teljes almondat állítása.

A *hogy*-os almondatnak két fajtája van. Az egyik esetben a főige tárgya egy mutató névmás (*azt, olyat…*), és az almondatban kifejezett állítás egészében a főmondat tárgya:

Olyat tett, hogy maga is megbánta,
Nem mondta, hogy fáj neki.

A *hogy*-os almondatnak egy másik fajtája, amikor a főmondat valamilyen szándékot, akaratot fejez ki, és az almondatban felszólító (kötő) módot használunk, így az almondat a főmondat célhatározójának tekinthető.

Azt kérdezte, hogy mi legyen vacsorára.
Úgy tett, hogy észre ne vegyék a csalást.

5.6.2 Almondat mint vonzat

A fenti példamondatok nem olyan szerkezetűek, mint az egyszerű vonatkozó névmás–mutató névmásos szerkezetek. Itt a *hogy* egy olyan kötőszó, amely egészében a főmondat mutató névmását helyettesíti ugyan, de tetszőleges állítással, nem pedig a két névmás által jelzett mondatrészt hozza közvetlen kapcsolatba.

Persze nem minden ige tárgya lehet ilyen *hogy*-os mellékmondat. Nem jók azok a mondatok, hogy

** Azt ette, hogy…* vagy ** Azt szigetelte, hogy…*

Emiatt a *hogy*-os mellékmondat is olyasmi, mint a vonzat. Az ige tulajdonsága, hogy lehet-e ilyen paramétere. Általában akkor van, ha valamilyen információközlést vagy információbirtoklást fejez ki:

Azt gondolta, hogy elég lesz.
Azt írta, hogy jól érzi magát.
Úgy értette, hogy te is ott akarsz lenni.

A másik esetben a főmondatban egy szándékot, akaratot kifejező segédige szerepel, de hiányzik az infinitívuszos főige. Ilyenkor a mellékmondat *hogy* kötőszóval kapcsolódva kibővítetten helyettesítheti a főigét.

Szeretném, hogy együtt nézzük át a leckét.	*Szeretném veled átnézetni a leckém.*
Kell, hogy legyen egy szabad hely.	*Kell szabad helynek lennie.*
A gyümölcs éretlen ahhoz, hogy leszedjék	*A gyümölcs éretlen még leszedni.*
Azért voltam a kertben, hogy kapáljak.	*A kertben voltam kapálni*
Képes vagyok arra, hogy megoldjam a feladatot.	*Képes vagyok megoldani a feladatot,*

Az utolsó példák is mutatják, hogy az infinitívuszi szerkezetek segédigés felfogásának van létjogosultsága. A legtöbb esetben ugyanúgy transzformálhatók *hogy*-os mellékmondattá, mint az egyéb segédigés mondatok. Az utolsó példamondatot ugyan lehet úgy is mondani, hogy *Képes vagyok a feladat megoldására.* Az almondattá transzformálás viszont sok egyéb vonzatnál nem lehetséges.

5.6.3 Mellérendelő összetétel

A mellérendelő mondatoknál két állítás szerepel, amelyek függetlenül elemezhetők. A kettő közötti viszonyt a kötőszó határozza meg. A tipikus mellérendelő kötőszavak ugyanazok, melyek a mondatrész-kötőszavak, de van egy-két kizárólag mondatkötőszó is. Lényeges helyesírási különbség, hogy akkor is vesszőt kell használni a tagmondatok közé, ha a kötőszó *és*, *vagy*, *meg*.

A mellérendelés a két mondatot azonos szintre helyezi, de lényeges, hogy a két mondat jelentése közti viszonyt a kötőszó vezérli.

A kóla hideg volt, <u>és</u> üdítőnek bizonyult.
A kóla, <u>habár</u> hideg volt, üdítőnek bizonyult.
A kóla hideg volt, <u>azért</u> üdítőnek bizonyult.
A kóla hideg volt, <u>de</u> üdítőnek bizonyult.

Az alapjelentés mindegyikben azonos mégis felismerjük a különbséget.

Elbeszélésben a mellérendelt mondatok gyakran jelentik események sorrendiségét is:

Mikor végre felébredt, kinyújtózkodott, megkereste a papucsát, és komótosan kiment a fürdőszobába.

5.6.4 Beágyazott mellékmondat

A magyar mondatszövést bonyolítja, emiatt az elemzést nehezíti, hogy egyik mondatba beágyazódhat egy másik. Értsd ezen azt, hogy egy mondat folyamát megtörheti egy almondat:

A levelet a postás, aki mindennap erre jár, nem dobta be a levélszekrénybe.
Péter, bár hideg volt, nem vett kabátot.

5.6.5 Amikor elmarad a vessző

Sajnos, a vessző nem csak akkor marad el az összetételek határán, ha helytelenül írjuk. Ha több vessző kellene egymáshoz közel, akkor nem kell mindegyiket kitenni. Például, ha az összetétel második tagja előtt egy mondatszó van, melyet az összetétel koordinálására használt mutató névmás vagy vonatkozó névmás vezet be, akkor a két vessző közül csak az egyiket kell kitenni. Hasonlóan, ha a beágyazott almondatot csak a koordináló mutató névmás előzi meg, gyakran elmarad a vonatkozó névmás előtt a vessző:

Amikor én akartam aludni, persze akkor kapcsolta be a szomszéd a rádiót.
Azt amit tegnap mondtam, nem kell komolyan venni

Az első mondat *persze* szavát lehet mondathatározónak is deklarálni, és akkor nincs baj a formalizmussal. A második eset viszont túl gyakran elkövetett hiba, ezért esetleg el lehet fogadni az elemzőnek.

5.7 Többértelműség a mondatok szintjén

Minden nyelvben természetes, hogy a szavak szintjén jelentkező többértelműség öröklődik a mondatelemzésben. Ennek egy része kiküszöbölődik a mondatelemzés során. Ha szigorúbb szintaxissal rendelkezik a nyelv, akkor van is rá esély, hogy nagy részét a mondatelemzés során egyértelműsíthetjük. Angolban ezért kisebb gond az, hogy egy szónak névszói és igei jelentése is van:

Flies are in the kitchen.
The plane flies high.

A magyar nyelvben a pozíció nem nagyon segít a szabad szórend miatt. Ha persze egy határozott névelő után egy olyan szó áll, amely főnév is és ige is lehet, egyértelmű a döntés:

A vár romokban hever.

Az angol elemzésénél nem szoktak ilyen részletes függőségi elemzést végezni. Megelégszenek, ha a mondat igéjét, alanyát, tárgyát, esetleg részeshatározóját nevesítik. A többi mondatrészt, névszói csoportot a prefixek úgyis meghatározzák. Ez többé-kevésbé igaz, mert olyan másodlagos vonzatszétszakadások nincsenek, mint a magyarban, ezért a névszói csoportok egybe maradnak. Sok olyan többértelműség fel sem merül, amely abból ered, hogy meg kell határozni, melyik rész mihez fűződjék. Persze így is akad. Két gyakran emlegetett példa a következő:

Time flies like an arrow
They saw the lady on the mountain with a telescope.

Több cikk jelent már meg taglalva, hány nyelvileg helyes értelmezése van ezeknek a mondatoknak. Az első mondat többértelműsége a szavak szintjén jelenik meg, melyet a szintaxis szerencsétlen módon nem tud feloldani. A különböző szóelemzések miatt így is, úgy is jó mondat kerekedik ki a végén, Ha magyarra fordítanánk, akkor az egybeírás és vessző használata miatt megszűnne a többértelműség.

Az idő száll, mint a nyíl. *Az időlegyek nyílvesszőt szeretnek.*

A másik mondat, ha nem tekintem az ige szószintű kétértelműségét (*saw* = *fűrészelik, látták*), akkor is a legtöbb általam ismert nyelven többértelmű, ha azt nézem, melyik névszói szerkezet hova kapcsolódhat.

Távcsővel látták a hegyen a hölgyet

Nem derül ki a mondat szerkezetéből, hogy a ki áll a *hegyen*, a néző vagy a *hölgy*, de az sem, kié a *távcső*, a nézőé, vagy a *hölgy* kezében volt. Esetleg a *hegy* tartozéka. Arról nem is beszélek, hogy eszköze vagy társa is lehet valaminek.

A magyar nyelv különös mértékben elviseli, ha egy mondatból akár lényeges rész is hiányzik. A hiány miatt elvileg az ige bármely vonzatkerete szerint elemezhetünk. Ha valamelyik vonzat hiányzik a mondatból, attól az még helyes. A szövegösszefüggésekből tisztázódhat a hiány. Más oldalról, ha egy másik vonzatkeret választunk, mint amit kéne, és emiatt egy határozószó vagy névszói kifejezés nem fér bele a vonzatkeretbe, akkor sok esetben szabad határozóként még értelmezhető lehet a vonzatkeretbe be nem férő rész.

5.8 Az elnyelődés

A magyarban időnként az okozhat gondot, ilyen értelemben többértelműséget, hogy egy elemzett rész több helyre is sorolható: több helyen is lehet része a szintaktikus szerkezetnek, ráadásul, ha mindkét funkciót betöltheti, nem feltétlen kell kétszer megjelennie. Két jellegzetes példa:

1. Ha birtokos szerkezet határozott névelővel kezdődik, a birtokos és a birtok is megkaphatja:

Az iskola diákjai ma sztrájkolnak

A határozott névelő lehet az *iskoláé*, de lehet a *diákok* névelője is. Ha megvizsgáljuk az ilyen eseteket, akkor a legtöbbször mindkettőé, de csak egyszer szabad kitenni. Megjegyzem, ha az elemző falánk módon a beágyazott struktúrához rendeli a névelőt, akkor mondatszinten nagy hibát nem követ el, mert az egész kifejezés határozott lesz, akár van névelője a külső kifejezésnek, akár nincs. A birtokviszony már azzá teszi.

2. A részes eset és az elváló, *-nak*-os bortokos:

Péternek elhoztam a kabátját.

Az ilyen mondatokban *Péternek* lehet a *hoz* ige nem kötelező vonzata, de lehet a *kabát* birtokosa. Ha viszont mindkettő, akkor sem szerepel kétszer a mondatban.

Hasonló eset a következő mondatban:

A téren játszadozó gyerekekkel találkoztam.

Itt a *téren* lehet a *találkozás* színhelye (szabad határozó), de lehet a játszadozó igenév szabad határozója. És ha mindkettő, akkor sem mondjuk kétszer.

5.9 Egy kísérleti mondatelemző

A korábbi megállapítások lényege, hogy vannak kötött és szabad szórendi szabályok. A 90-es évek végén megkíséreltünk egy olyan mondatelemzőt létrehozni, amely egy elvi ember-gép kapcsolat része lehetett volna. Mivel az ember-gép kapcsolatnál a lényeg, hogy a mondat szavai közti összefüggést tárjuk fel, ráadásul olyan formalizmussal, melynek eredményét egy esetleges szemantikai elemző felhasználhatja, célszerű volt egy vonzatgrammatikára építeni. Tehát nem elég, ha megmondjuk, hogy helyes-e a mondat, hanem a szavak közötti összefüggés szerkezetére is fényt derítünk. Ez a szerkezet nem szemantikai, jelentést nem hordoz, de a szintaktikai szerkezetek mögött szemantikai háttér is lapul. A jelentések függvényében a szavak más, más vonzatkerettel rendelkeznek. Ha felismerjük a szavak aktuális vonzatkeretét, akkor a jelentéshez is közelebb kerülünk.

5.9.1 Az algoritmus lényege

Az elemzés szósorrendi és szórendfüggetlen szabályokra épül. A szórendi szabályokon kívül – főként a vonzatok fellelése, a szabad határozók felismerése miatt a szokásostól eltérő hatékony algoritmust használtunk. A többértelműséget súlyozási kiértékeléssel csökkentettük, melyek segítségével már menet közben elvethettünk sok alternatívát:

- Az elemző egy a **4.12.1**-ben leírt szótani elemző eredményére épít. Ha már a szöveget némi heurisztikai módszerrel mondatokra daraboltuk, akkor elemezzük a szavakat. Az elemzett szóalakok listája olyan elemekből áll, melyek tartalmazzák a szóalakot, a szó tövét, a képzett alak szófaját, a ragok és jelek kódját és a használt képzők sorát. Ez utóbbit a vonzattranszformáció kezelése érdekében. Természetesen alternatívákkal együtt, hisz a morfológiai elemzés nem egyértelmű. Ehhez társítjuk a szó vonzatkereteinek listáját, melyet egy kiegészítő vonzatszótárból vesszük.
- A mondatnak ezek után megkeressük a fejét, ami vagy egy ragozott ige, vagy feltételezzük, hogy névszói mondat. Ha a mondatban van olyan szó, amelynek csak igei elemzése van, akkor nem próbálkozunk névszói elemzéssel. Hasonlóan, ha nem találunk ragozott igét, akkor feltételezzük, hogy névszói mondatról van szó.
- Feltételezve, hogy igei a mondat, összehalásszuk az igei kifejezést: az igét, segédige vagy szenvedő szerkezet esetén a fő igét is, az igekötőt. A vonzatszótárból megállapítjuk, mik lehetnek a lehetséges vonzatkeretek, természetesen az esetleges vonzattranszformációt figyelembe véve. Ezek után a mondatban megkíséreljük a névszói kifejezéseket meghatározni, kihagyva a már fellelt igei szerkezetet.
- Az elemzett lista alapján megkíséreljük a lehetséges névszói kifejezéseket behatárolni. A mondatot jobbról balra elemezzük. Ennek az az oka, hogy egy kifejezésnek általában a legjobb oldali szava határozza meg, hogy milyen vonzatot elégít ki, így valószínűleg egy legfelső szintű névszói kifejezéssel találkozunk először. Az elemzés során a beágyazott kifejezéseket automatikusan beolvasztjuk a talált kifejezésbe.
- Ha a soron következő szóformának van olyan elemzése, ami kielégít egy vonzatot (pl. megfelelő raggal rendelkező szó, megfelelő névutót vagy határozószót találunk), akkor ebből a szóból kiindulva

megkeressük a hozzátartozó névszói kifejezést. Ezt felfűzzük a vonzatkeretlista azon alternatíváiba, ahol szerepel a vonzat. A továbbiakban ezeket az alternatívákat vizsgáljuk először.

- A névszói kifejezést falánk, direkt módszerrel elemezzük. Ez azt jelenti, hogy ha a névszói szerkezet szintaxisának megfelelő szófajú (és ragtalan) szót találunk, vagyis jobbról balra haladva bővíthető a következő szóval az eddig felderített névszói kifejezés, akkor azt be is építjük a kifejezésbe. Ha az éppen beépített szónak vannak vonzatai, és a következő (tőle balra lévő) rag, névutó vagy szó a vonzatnak megfelel, vagyis van arra esély, hogy egy olyan kifejezés végére leltünk, mely a vonzatot kielégíti, akkor a beágyazott kifejezést rögtön elkezdjük keresni, és ha megtaláljunk az elejét is, hozzácsatoljuk az megtalált vonzatot az őt követő szóhoz.
  - Halmozott névszói kifejezésnél az összes alkotórésznek ugyanazt a vonzatot kell kielégítenie. Ha az egyik alternatíva vonzata bővebb, a másiké szűkebb (pl.: HOVA és *-ra/re*), akkor a szűkebb vonzatot tartalmazó alternatíva kieshet, ha különböző ragok vagy névutók vannak összekapcsolva
  - Közben megjelöljük azokat a helyeket, ahol esetleg széteshet a névszói kifejezés két független részre.
- Az így megtalált vonzatok alapján kiértékeljük a vonzatkeretek kitöltöttségét. A különböző vonzatok és vonzatkereteknek más a súlya. Az összetettebb vonzatkeretek többet érnek, mint a rövidebbek. A kulcsszavas vonzatok jobban stimmelnek, mint a ragos, névutós szerkezetek. Ha egy kötelező vonzat hiányzik – nem találtunk az algoritmus alapján megfelelő elemet, az jobban csökkenti a stimmelés valószínűségét, mintha csak egy opcionális vonzat hiányozna… stb.
  - A jobban értékelt vonzatok mellett a maradékot, fel nem használt mondatrészeket megkíséreljük másodlagos vonzatként és szabad határozóként értelmezni.
  - Ha a vonzat mellékmondat, vagy egyébként a kötőszavak, vesszők alapján összetett mondatot érzékelünk, a vesszőkkel leválasztott részt tovább elemezzük.
- Az elemzést, ha kell, névszói mondatként is elemezzük, hasonlóan kiértékeljük.
- Számba vesszük a kihagyott részeket, és ezeket a negatív súllyal értékeli számba az algoritmus.
- A különböző elemzéseket összevetjük, és csak egy bizonyos szintet elérőket adjuk tovább, mint lehetséges alternatíva.

5.9.2 A vonzatszótárról

A vonzatok egyszerű szöveges állománya könnyen módosítható. Ami lényeges, hogy az igekötős igék vonzatait külön felvettük, és az igekötő nélküli igéknél, mint vonzat, szerepel az igekötő, ami azt jelenti, hogy az elemző, ha talál egy megfelelő elváló igekötőt a mondatban, akkor azt a vonzatát keresse az igének amely az igekötős bejegyzésnél szerepel:

```
ad+2000ki,+2000meg,+2000össze,+2000el,+2000fel,+2000vissza,+2000be,+11T+71NEK(+51ÉRT+430
NI),+100lehetôséget+31RE(+71NEK).
bead+11T+902HOVA.
kiad+11T+71NEK,+11T+31RE.
elad+11T+71NEK+902HOVA.
felad+11T,+11T(+902HOVA+71NAK).
kiad+11T+71NEK,+11T+31RE.
megad+11T+71NEK,+11T(+70VEL).
összead+211T,+11T+70VEL.
visszaad+11T(+71NEK+903HONNAN).
```

A forma önmagáért beszél. Nem csak igéknek, de névszóknak és névutóknak is itt szerepel a vonzata:

```
kapcsolat+70VEL.
kapcsolattartás+70VEL.
kategóriájú+3000MELY.
keresztül+30N.
képes+31RE
```

5.9.3 Példa

: *A válás 8 éve történt*

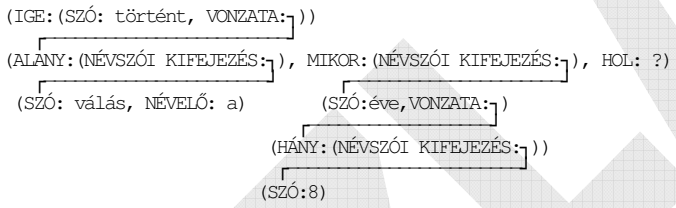
A mondatszegmentálás és morfológiai elemzés után:

```
((A, (a:határozott névelő)),
 (válás, (vál[ik]:ige, ás:képző)),
 (8, (8:számnév)),
 (éve, (esz[ik]:ige, ve:határozói igenév képző),
       (év:főnév, e:birtokjel),
       (éve:időhatározószó)),
 (történt,(történ[ik]:ige, t:igerag),
       (történ[ik]:ige, t: melléknévi igenév képző)),
 (.:írásjel))
```

Belső kódokkal:

```
A:
  a|3804
 válás:
  vál|10005,8004
 8:
  8|23004
 éve:
  éve|1410
  év|21004,2013
  esz|10005,1530
 történt:
  történ|10005,8002
  történ|10005,1103
 .:30046
```

Elemzése:



Szöveges kimenetként:

```
*ragozottIGE*:
  ([történt:+10005+ 1103],
   *Vonzata*:
    (*Vonzattípus*:(ALANY), *Vonzott*:
      (*Névszói*:
        ([válás:+21020],
         [A:+3804])),
      *Vonzattípus*:( 1904,), *Vonzott*:
       (*Névszói*:
          (*Vonzattípus*:( 3300,), *Vonzott*:
            (*Névszói*:
             ([8:+23004])))))),
      *Vonzattípus*:( 1901,)))
```

5.10 Szabad-e a szabad szórend?

A magyar szabad szórendű nyelvnek számít. Valóban, a következő háromszavas mondat szavait tetszőleges sorrendbe tehetjük, a jelentés lényegében nem változik:

Péter szereti Katit. Péter Katit szereti. Katit Péter szereti.
Katit szereti Péter. Szereti Katit Péter. Szereti Péter Katit.

Ha viszont a következő nem teljes mondat (névszói kifejezés) szavait semmi esetre sem cserélhetjük meg:

három sárga csőrű kiskacsáról

Ezek szerint, nyelvükben vannak szórendfüggő és szórendfüggetlen szabályok. A nyelvekben egy általános szabály, hogy azok a szavak, mondatrészek melyek közti kapcsolatot nem jelöljük semmilyen ragozással, prefixszel, toldalékkal, azok között a kapcsolatot a szórend adja meg. Ahol viszont a kapcsolat milyensége morfológiai elemekkel jelölt, ott a szavak, mondatrészek „elmászhatnak” egymástól. Magyarban, a főnévhez tartozó melléknév mindig megelőzi a főnevet, mert a jelző szótári alakban szerepel. Oroszban viszont a melléknevet nemben, számban, esetben egyeztetni kell a főnévhez, ezért sorrendjük szabadabb, mint a magyarban (bár általában náluk is megelőzi a főnevet).

Összefüggő kötött sorrendű mondatszerkezet a névszói kifejezés, és kötött szórendű szabályok találhatók az igei szerkezetekben is. De ezekről korábban írtam. Az viszont igaz, hogy ha valamilyen szempontból szabadság van, mit hogyan fejezünk ki egy nyelvben, akkor a szabadság sohasem teljes. Valamilyen,

modellünkben nem kifejezhető különbség mindig akad. Nos, a magyar nyelvben a szósorrend is hordoz információt. Ez az információ inkább pragmatika jellegű. Pontosabban azt lehet felismerni a mondatban, mi az információ közlőjének szándéka, mi az amit, mint új információt közölni szándékozik. A mondat lényegi tartalmát nem változtatja, de ennek ellenére nagyon fontos felismerni, illetve mondatszerkesztésünkben követni az ennek megfelelő szabályokat.

5.10.1 Téma – réma – megjegyzés: kötések a szabad szórendben

A magyarban a függőségi viszonyokon kívül más szerkezet is található. Ez az a szerkezet, mely alapján meghatározható, mi a lényeges új információ a mondatban, milyen korábbi közléshez akarjuk csatolni, és mi az, ami csak kiegészítő, ilyen értelemben nem lényeges. E három rész, a réma (focus), a téma (topic) és a megjegyzés (comment). Ha egy külföldit hallunk, aki jól elsajátítja nyelvünket, ezt nem szokta figyelembe venni, ehelyett saját nyelvének szórendjét adja vissza a magyarban is. Ha megnézzük a korábban említett háromszavas mondatot, akkor az adott szósorrenddel a hangsúly, vagyis a lényeges információ vagy az ige előtti szón van, vagy az ige hordozza:

Péter szereti Máriát. *Péter szereti Máriát.*
Péter Máriát szereti. *Péter Máriát szereti.*
Máriát Péter szereti. *Máriát Péter szereti.*
Máriát szereti Péter. *Máriát szereti Péter.*
Szereti Máriát Péter.
Szereti Péter Máriát.

Megállapíthatjuk, hogy az új információt bevezetheti a téma, és követheti a megjegyzés.
A kötelező sorrend:

téma, réma megjegyzés
topic, focus, comment

Ezek közül a réma helye meghatározható. Vagy maga a ragozott ige, vagy a közvetlen előtte álló rész. A többi ehhez relatív helyezkedik el.

A **téma**: a bevezetés, mely elhelyezi a mondandó alapkörülményeit, mintegy bevezeti, miről lesz szó. Témából több is lehet egy egyszerű mondatban, de ritka a kettőnél több. Gyakran utalás helyre, időre, körülményre, a szövegben korábban elhangzottra, vagy az általánosan ismert dolgokra. Beszédben a végén felfelé ível a hanglejtés, utána rövid szünetet tartunk.

A **réma** a közölni való információ, az új adat, kérdő mondat esetén a kérdés tárgya. Ebből egy egyszerű mondatban csak egy lehet. Azt viszont nem mindig lehet eldönteni, hogy az egész névszói szerkezet, vagy csak egy része a központ, az új információ. Gyakran csak egy tagadószó. A beszédben a hangsúly gyakran eldönti ezt is. Az egzisztenciális mondatokban, ha más nyelvre kell fordítani, ezt a részt érdemes állítmánynak tekinteni, így megtalálhatjuk, mi lehet pl. igés angol mondatra fordításnál az alany.

A **megjegyzés** további kiegészítőket tartalmazhat. A végén leszálló a kiejtés dallama.

Mivel a nyelv funkciója a közlés, réma nélkül nincs mondat. A téma, és a megjegyzés elmaradhat. A réma kapja beszédben a fő hangsúlyt, de a szórend is utal a kiosztásra. A magyar nyelvben az ige (névszói mondat esetén a létige helye) adja meg a réma helyét. Vagy maga az ige a fő közlendő, vagy a ragozott ige előtt áll a réma. Ha van, megelőzi az egy-két téma, és a megjegyzés hangsúlytalanul követi az igét.

A fenti példamondatokból is látszik, hogy írott szöveget többféleképpen is lehet felosztani témára, rémára. A példamondat utolsó két szórendjében, mivel az ige előtt nem áll semmi, csak az ige lehet a lényeget hordozó rész, míg a korábbiakat hangsúlytól függően legalább kétféleképp lehet felbontani. Egyértelműsíthet viszont néhány olyan jel, mely kizárhatja vagy kötelezővé teszi a téma pozícióját.

Az igekötős igéről, ha megelőzi a réma, akkor elválik az igekötő, helyet adva a fókusznak. Ha nem válik le az igekötő, akkor csak az ige lehet a fókuszban. Egy kivétel, az igemódosító vagy tagadószó, amely, ha nem maga a fókusz, beékelődhet az igekötő és az ige közé akkor is, ha az ige a fókusz:

Moziba el nem megyek

Ez persze akkor is lehetséges szósorrend, ha az igekötő a téma, és a tagadószó a réma:

Moziba el nem megyek

Ha viszont a tagadószó a réma és az igekötő nem téma:

Moziba nem megyek el.

Kérdő mondatokban a kérdőszó vagy névszói kifejezés az ige előtt áll. A tagadás, a mellékmondatra utaló mutató illetve vonatkozó névmás is leggyakrabban a fókuszban áll.

Ki ölte meg Kennedyt?
A kérdésre tudja-e a választ valaki?
Senki se tudja a választ?

A téma-réma összefüggés ad választ arra, miért szakadnak el összetartozó mondatrészek, legalábbis akkor, ha elválaszthatók, tehát nincs kötelező szósorrend, mert a toldalékok alapján koordinálható a kapcsolat. Ha egy leválható rész a téma vagy a réma, akkor valahova az ige elé kell kerülnie, akkor is, ha a többi rész például a megjegyzés részben szerepel. Ilyenek a *-nak*-os birtokosok, a másodlagos vonzatok…

Péternek bedöglött a kocsija.
Péternek döglött be a kocsija.
A kocsija Péternek döglött be.
A kocsija Péternek bedöglött.

Érdekes szituáció, ha igét kell a témába helyezni. Fizikailag csak úgy lehet az igét önmaga elé helyezni, ha megduplázzuk egy főnévi igenév formájában:

Szeretni szereti Máriát.

Igekötős igékkel alternatív megoldás, ha önmaga helyett csak az igekötőt küldi a fókusz elé az ige:

Megismerni Péter ismerte meg Máriát, nem Mária Pétert.
Meg Péter ismerte Máriát, nem Mária Pétert.

Nem minden nyelvben ilyen a téma-réma sorrend. Az angol például ha nem élünk a hangsúly, a prozódia eszközével, a mondatok, kifejezések végére koncentrálják a lényeget. Ezért írásban, ha egy mondat alanyát akarják a fókuszba emelni, akkor ezt szenvedő szerkezettel oldják meg, így a *by* szócskával a mondat végére kerülhet a lényeg, az alany:

Mary is loved by Peter

vagy egy összetett mondatban az első tag utolsó elemként szerepel a következő mondat alanya:

It is Peter who loves Mary

Az első típusú alárendelő összetett mondatban gyakran a főmondat mutató névmása a témában vagy a rémába helyezkedik el, az almondat vonatkozó névmása viszont szinte kizárólag a téma része, sohasem lehet a ragozott ige után. A kérdő mondatnál a kérdőszó természeténél fogva a fókusz (réma). Ilyenkor persze a mutató névmás nem lehet a fókuszban.

Mikor fogják előhúzni azt a kartotékot, mi jogom sérti meg.

A beszélt nyelvben a réma mindig hangsúlyt kap, de ez az írásban nem látszik. Ez ad alkalmat arra, hogy beszélt nyelven eltérjünk ettől a szórendtől. Előfordul, hogy a mondat végére kerül a hangsúly, így a réma, de ez nem túl szép. Talán akkor fogadható el, ha az alárendelt összetétel főmondatában a mutató névmás hátra kerül, hogy közelebb legyen a hozzá tartozó vonatkozó névmáshoz:

Tegnap vásároltam azt, amit tegnap előtt elfelejtettem.
Tegnap vásároltam cukrot, kenyeret, vajat és paprikát.

És mint a példákból is jól látszik, az alárendelés koordinátoraiból a vonatkozó névmás a téma része.

6. Mondatelemzés környezetfüggetlen nyelvtannal

Noam Chomsky az ötvenes években, amikor első átütő munkáit publikálta, arról írt, hogy a természetes nyelvek, nyelvtanok ugyan különböznek, de valamilyen közös metarendszerre épülnek. Egy ilyen metarendszernek gondolta a környezetfüggetlen nyelvtant. Azóta sok idő telt el, de a mai gépi nyelvészek többsége azóta is környezetfüggetlen nyelvtannal, vagy ahhoz hasonlóval igyekeznek leírni a nyelveket. Ez az eszköz valóban hatékonynak bizonyult, egyrészt a nyelvek leírása szempontjából, de az implementáció szempontjából is.

A leírás szempontjából azért hatékony, mert jó eszköz nyelvi fogalmak, elsősorban szintagmák absztrakt használatára, másrészt a gyakori sorrendfüggések és egymásba ágyazódások leírására.

Az megvalósítás szempontjából azért, mert – bár elsősorban a programnyelvek implementálásának tapasztalatára építve – sok olyan általánosan elérhető hatékonyan használható módszert, eszközt fejlesztettek ki, melyek a környezetfüggetlen nyelvtanok elemzésére, fordítására alkalmasak. Általános értelemben generatív nyelvtanok között az általános grammatikák (Chomsky 0. osztály) alapján nem lehet becsülni az elemzés idejét. A környezetfüggő (1. osztály) nyelv mondatainak elemzési ideje exponenciális a bemenet függvényé-

ben. A környezetfüggetlen nyelvek (2. osztály) esetében az általános elemzők időigénye köbös az input hosszával. A reguláris nyelvek esetén lehet csak elérni a lineáris nehézséget. Pontosabban, a determinisztikus környezetfüggetlen nyelvek is csak lineáris nehézséget jelentenek, de az élő nyelveknél erre nincs remény. Ha egy nyelvtan, a hozzá tartozó nyelvvel nem elemezhető egyértelműen, akkor determinisztikus elemzésről szó sem lehet. Marad a köbös nehézség és környezetfüggetlen nyelvtanok használata. Felmerülhet esetleg reguláris nyelvtanok alkalmazása is, de a nyelvészek nehézkesnek találják a nyelvi jelenségek leírására. Azért ne vessük teljesen el ezt a lehetőséget.

E két előny mellett eltörpülnek azok az indokok, hogy a nyelvi jelenségek nem feltétlen környezetfüggetlen nyelvtannal leírható dolgok, másrészt vannak nem generatív módszerek, melyek időnként hatékonyabbak a generatív módszereknél.

6.1 Minta angol mondatok leírására

Angolra, ahol dominálnak a sorrendi szabályok, aránylag könnyű olyan szintaxist írni, ahol a vonzatszerkezet ugyan nem látszik, de a fő mondatalkotórészek megtalálhatók, a kifejezések szerkezetei felismerhetők, és tulajdonképpen a nyelvileg helyes mondatok felismerhetőek, alapszerkezetük megfejthetők.

Négy mondattípust különböztetek meg. Kijelentő, kérdő, felszólító, címmondatot és összetett mondatot.

```
sentence-> statement | question | order | headline | sentence, coordinator, SVCOP.
```

A kijelentő mondat részeinek sorrendje: elől van az alany, amit csak néhány bevezető határozói kifejezés előzhet meg, és opcionálisan ponttal végződik.

```
Statement → SVCOP, ["."].
SVCOP → [circumstance], ",", subject, VCOP.
```

A maradék rész attól függően, hogy tranzitív-e az ige, követhet 0, 1 vagy 2 tárgy, és az egyéb nem alanyesetű részek, vagy egy létige, melyet követhet egy predikátum:

```
VCOP → verbP, [[object,] object], [circumstance] |
       verbP, object, ["to", object,] [circumstance] |
       verbbe, [predicate], [circumstance].
```

Az alany és a tárgy hasonló szerkezet:

```
Subject → nounP.
Object → nounP.
```

A predikátum hasonlóan egyszerű:

```
Predicate → adjectiveP |
            nounP |
            circumstance |
            participle, [[object], object], [circumstance] |
            Participle, object, "to", object, [circumstance].
```

Az egyebek sok minden lehet határozótól almondatig.

```
Circumstance → "here" | "there" | "everywhere" |
               modifier | subordinator, SVCOP | localisator, SVCOP.
Modifier → PREP, noun | PREP, number.
Subordinator → "if" | "since" | "because" | "although".
Localisator →"when" | "where" | "why".
```

A kulcs az igei rész:

```
verbP → [adverb], verbform |
         verbbe , participle |
         semiauxv, "to", infinitive |
         auxv, infinitive |
         verbhave, participle |
         verbhave, [adverb], "been",
         participle.
         verbbe, participle |
         verbhave, [adverb], "been", participle.
```

Az igék persze tovább bonyolódnak:

```
Verbbe → be, ["not"] |
         semiauxv, ["not"],"to", "being" |
         auxv, ["not"], [semiauxvINFI, "to"], "being" |
         "having", "been".
```

Lehetne még bonyolítani. A leírásban főként szótári elemek hiányoznak, no meg az, hogy például az ige és az alany száma, személye egyezzen, számnév után többes számot használjunk, csak tranzitív igének lehet tárgya… Ez utóbbiakat a szabályok sokszorozásával lehet megtenni. A felszólító mondat, a címmondat egyszerű szerkezetű, de a kérdő mondatnál valamivel többet kell írni. A szótári tételeken kívül így is meg-

úszhatjuk párezer szabállyal. Ha kevesebb írásmóddal akarjuk ugyanezt a szintaxist elérni, akkor más módszerhez kell folyamodni. Lásd a ragnyelvtanokat.

Azért lehetett ilyen egyszerű szintaxist írni, mert a vonzat jellegű összefüggésekkel nem törődtem. Ami ebből a leginkább hiányzik, de könnyen pótolható, az a phrasal verb használatánál az elmászó igerész. Ez hasonló jelenség, mint a magyar igekötők esete:

Take off your old coat!
Take your old coat off!

Ez viszont könnyen megoldható környezetfüggetlen nyelvtannal, de csak akkor fontos, ha a vonzatokat is kezeljük.

6.2 Ami nem fér bele a környezetfüggetlen nyelvekbe

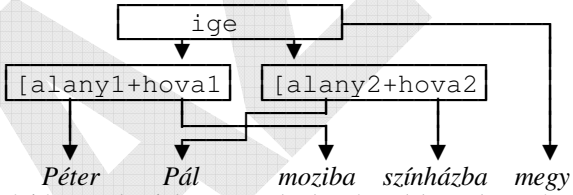
A nyelvi formák szabályok közt vannak olyan jelenségek, melyeket nem lehet CF nyelvtannal megadni. Általában azt lehet kimutatni, hogy a jó struktúrához nem lehet olyan levezetési fát adni, amelyikben a levezetési ágak ne kereszteznék egymást. A nyelvészek ezt úgy hívják, hogy nem konfigurálhatók. Ha ilyen szerkezet nincs, akkor a nyelv konfigurálható. Persze, lehetséges, hogy a megadott nyelvtan nem jó, más nyelvtannal kijön a konfigurálhatóság. Nos, van olyan jelenség, amit nem lehet elvileg sem megadni CF grammatikával. Tekintsük a következő mondatot:

Péter, Pál, Mari rendre moziba, színházba, úszni ment a hétvégén.

Ez azzal ekvivalens, hogy

Péter moziba, Pál színházba, Mari úszni ment a hétvégén.

Ha a levezetési fában fel akarjuk tüntetni, hogy ki, hova ment, akkor második mondatnál ezt természetes módon megtehetjük, míg az első formában a leírófa ágai keresztezik egymást. Ezt, ha figyelembe vesszük, hogy bárhány személy bárhány cselekvését kéne párba szedni, de a sorrend a megfelelő legyen ekvivalens azzal az absztrakt problémával, van-e olyan CF nyelvtan, amely azokat a karakterláncokat fogadja el, amely egy sztring ismétlése. $\{ww \mid w \in \Sigma^*\}$ Erre a közismert válasz, hogy nincs, Ha korlátoznánk a felsorolt elemeket, akkor ugyan kitalálhatnánk olyan szintaxist, ami elfogadja a helyes mondatot, de a szintaktikus levezetés nem tükrözné a páronkénti összefüggést. Már két elemnél is a fellépne az a helyzet, hogy a levezetési fa két ága keresztezi egymást, ha csak lefelé mutató élek lehetnek.



Hasonló konfigurációs gondot jelent a szabad szórend következményeként elmászó másodlagos vonzatok, igekötők, birtokosok felfűzése a levezetési fában, ha azok lokálisan nem ott vannak, ahova tartoznak.

További gond az elnyelődés, amikor ráadásul nem is faszerkezetű, hanem háló jellegű a levezetési struktúra: nem csak szétágazás, hanem csatlakozás van a függőségi struktúrában.

6.3 Ami sok a környezetfüggetlen nyelvtannál

Az elméleti megvalósíthatatlanságon kívül további gond, hogyha hagyományos CF nyelvtannal szeretnénk leírni a nyelvtant, a szabályok száma irdatlan nagy lenne, Ha a nem a teljes angolt nézzük, akkor is sok, már pedig az egyeztetések megsokszorozzák a szabályok számát. Csak az ige-alany egyeztetés hatszorosára növeli a szabályokat, de a sorrendfüggetlenség – ha a vonzatkeretet is meg akarjuk kapni – nagyságrendekkel növeli a szükséges hagyományos környezetfüggetlen szabályok számát. Emiatt egyes formalizmusok megengedik a szabadszórendű szabályok használatát, melyet úgy kell értelmezni, hogy a szabályok jobb oldalán levő elemek tetszőleges permutációja megengedett. Ezzel ugyan a függőségek nem követhetők, de a helyesség ellenőrizhető.

A sorrendfüggetlen függőségi leírás és az egyeztetés követése leírásának hatékony eszköze az úgynevezett többszintű nyelvtan.

7. Többszintű nyelvtanok. Rag- és attribútumnyelvtanok

A többszintű nyelvtanok használata a programnyelveknél vált gyakorlattá. Az alapvető gond ott az volt, hogy a 60-as évek elején beváltnak hitt környezetfüggetlen nyelvtan nem volt elegendő a programnyelveknél arra, hogy a típusellenőrzést is a formalizmussal megadják. Ha egy változó valaminek deklarálva volt, akkor egy ettől teljesen független helyen a fordítási időben ellenőrzik, megfelelő-e a változó használata. Ez ugyan igaz, de a megoldásra nem adtak formális leírást, inkább szövegesen adták meg a szabályokat, melyet ügyes programozók aztán meg is valósítottak a fordítójukban.

A 60-as évek végéig kialakult ennek is a megoldása, a kétszintű nyelvtan, melynek lényege, hogy a nyelvtant is formálisan írják le, mely leírás alapján nem hagyományos környezetfüggetlen nyelvtant kapunk. A nyelvtan alapelemeit a metanyelvtan határozza meg. Így lehetséges, hogy az így keletkezett fő nyelvtan grammatikai, esetleg terminális jeleinek száma végtelen. Egy adott program elemzésénél ennek csak véges részhalmazát kell használni, ezért a környezetfüggetlen nyelvek elemzésénél bevált módszerek mégis alkalmazhatók voltak. A hatvanas években ilyen általános fordítónyelv volt a CDL (compiler definition language) melynek segítségével implementálták a kellően bonyolult Algol68 nyelvet is.

Általános értelemben a kétszintű nyelv azt jelenti, hogy egy metanyelvtan segítségével definiáljuk egy másik nyelv leírását, szintaxisát. Ha ezt nézzük, így elvileg bármilyen bonyolult generatív nyelvet meg lehet adni akár környezetfüggetlen metanyelv segítségével. Gondoljunk arra, hogy egy Túrig-gép CF grammatikával definiálható.

A gyakorlati alkalmazásoknál nincs teljes szabadsága a metanyelvnek. Az egyik egyszerűsített modell szerint a metanyelv segítségével a fő nyelv terminálisait lehet „deklarálni”, a másik szerint pedig a grammatikai jeleit is. Ezzel lehet feloldani a típusellenőrzést már fordítási időben. A nagy különbség abból adódik, hogy így a hagyományos generatív nyelvekkel szemben az ábécé nem véges, hanem a metanyelv által meghatározott, akár végtelen elemből is állhat.

Ha csak a terminálisok meghatározására kell a metanyelv, akkor ha a M-mel jelöljük a metanyelvtant, L_M-mel az ez által generált nyelvet. Ez az eredeti $\sum$ ábécére épül, mely véges halmaz. Az igazi (második szintű) nyelvnek, bár lehet, hogy környezetfüggetlen sémákat tartalmaz, az ábécéje az L_M, ami nem feltétlen véges, emiatt nem feltétlen környezetfüggetlen az a nyelv, amely így két lépcsőben kialakul és eredendően $\sum$ az ábécéje. Ennek ellenére az elemzése épülhet sok olyan módszerre, egyszerű veremtechnikára, amelyeket CF nyelvek elemzésére dolgoztak ki.

A két szint által megadott nyelvtan sokszor implementálható úgynevezett kétlépcsős fordítással. Ha e két lépcső mindegyike veremfordítóval hajtható végre, akkor a kapott nyelv előállítható két környezetfüggetlen nyelv metszeteként, és megfordítva, ha egy nyelv előállítható két környezetfüggetlen nyelv metszeteként, akkor elemezhető egy veremfordító és egy veremautomata egymásutánjával. Ha bármelyik szint véges automatával/fordítóval megvalósítható, akkor a kétszintű nyelv megmarad környezetfüggetlennek.

A két szintnek kettős jelentősége van.

1. Megoldható a típusellenőrzés – élő nyelveknél a függőség és az egyeztetés
2. Tömöríthető, egyszerűsíthető a leírás

A tömörítésnek egy szép példája a rag- illetve tulajdonságnyelvtan (affix, attributum grammar). A ragnyelvtan lényege, hogy a nem terminálisokat ruházzuk fel olyan plusz információval, melyeket aztán a levezetési szabályokban használhatunk egyeztetési, illetve öröklődési jelenségek megadására. A metanyelv határozza meg, hogy a második szint nem terminálisainak milyen tulajdonságai lehetnek, melyet a második szint levezetési szabályaiban felhasználhatunk.

A **6.1**-ben említett nyelvtant például kiegészíthetjük olyan ragokkal, mint `num` és `pers`, melyet az igei rész és az alanyi rész is visel.

Tehát a metanyelvben deklaráljuk a ragokat – ez által a második szintű nyelvtan grammatikai jeleit

```
num → sing | plur
pers → 1 | 2  | 3
```

és ezek után a fő nyelvtan alkalmas része úgy alakulhat, hogy:

```
statement→ SVCOP, ["."].
SVCOP → [circumstance], ",", subject(num, pers), VCOP(num, pers).
```

Itt egyeztetés szerepelt

```
VCOP(num, pers) → verbP(num, pers), [[object,] object], [circumstance] |
              verbP(num, pers), object, ["to", object,] [circumstance] |
              verbbe(num, pers), [predicate], [circumstance].
```

Itt pedig öröklődés szerepelt

A formalizmus megint önmagáért beszél. A nyelvtani leírás hasonló a CF-nél megszokotthoz, azzal a kivétellel, hogy a szokásos nem terminálisok ki vannak egészítve tulajdonságokkal, ragokkal. A példánkban jól meghatározott, hogy milyen grammatikai szimbólum milyen típusú, tehát milyen ragokat vehet fel. Egy nagyon hatékony eszköz lehet, ha a metanyelv reguláris, esetleg véges. Ilyenkor az eredménynyelv is megmarad környezetfüggetlennek. Erre láthatunk példát a **7.1**-ben.

Lehet ettől összetettebb a metanyelv. A másik véglet, ha csak a tulajdonságok maradnak meg, Például kulcsszavas ragazonosítók után azok értékeit adjuk meg, és már nincs helye a főnyelvtanban másnak, mint ilyen tulajdonságleírások közti levezetésnek:

```
(CATEG=statement) → (CATEG=SVCOP)(CATEG=fullpoit, suface=".")
(CATEG=SVCOP) → (CATEG=subject,num,pers)(CATEG=VCOP,num,pers) |
   (CATEG=circumstance)(CATEG=comma,suface=",")(CATEG=subject,num,pers) (CATEG=VCOP,
   num, pers).
(CATEG=VCOP,num,pers)→(CATEG=verbP,num,pers)(CATEG=object,num,pers)…
```

Bár a leírás kevésbé olvasmányos, mégis sok helyen alkalmazzák, mert teljesen általános. Ha a metanyelv megengedi, az attribútumok akár egymásba ágyazottak is lehetnek. Nincs semmi sem megkötve. A szabályok nyelvtana, tehát a metanyelvtan, CF, ami erre épül, az pedig CF jellegű,

7.1 Fejközpontú unifikációs nyelvtan (HPSG)

http://xsb.sourceforge.net/manual1/node114.html

http://hpsg.stanford.edu/

http://ling.hung.u-szeged.hu/szecsenyi/hpsg.hu.pdf

7.2 Logikai alapú megoldás (DCG)

7.3 Egy egyszerű, de hatékony nyelvtan természetes nyelvekre (AGFL)

Ragnyelvtanra nagyon jó példa C. H. Koster által megalkotott véges háló feletti ragnyelvtan. A természetes szövegek elemzésére készült általános eszköz egy kétszintű nyelvtant hatékonyan implementáló univerzális parszer. A szintaxis metanyelvi részében deklarálódik, milyen ragtípusok vannak,

Példaként lássuk a korábban felírt angol nyelvtan pontosítását.

A metanyelvben deklaráljuk, milyen ragosztályok milyen halmazt jelentenek:

```
#-------------------------META SYNTAX-------------------------#
NUMB:: sing | plur.
PERS:: first | secnd | third.
CASE:: nom | gen | dat | acc.
TRAX:: trav | ditr.
TRAN:: TRAX | intr.
PREP:: none | to | from | at.
TENS:: infi | prpl | papl.
TIME:: prst | past.
COMP:: abso | comp | supl.
```

Erre építve használjuk azt a nyelvtant, melyben a fenti ragokkal bővülnek a grammatikai szimbólumok:

```
sentence: question;
          statement;
          order;
          SVCO phrase,coordinator,sentence.
statement: SVCO phrase,["."].
question: vSvCO phrase,["?"];
          localisator,vSvCO phrase,["?"];
          object question,vSvCnO phrase,["?"];
          PREPOS(PREP),object question,vSvOnC phrase(PREP),["?"];
          xSvCO phrase,["?"].
order: VCO phrase(plur,secnd),["!"].
coordinator: "and"; "or"; ",".
```

```
localisator: "when"; "where"; "why".
object question:"what",[object without determiner];
                "which",object without determiner;
                "whose",object without determiner.
circumstance: "here"; "there"; "everywhere";
              modifier;
              subordinator,SVCO phrase;
              localisator,SVCO phrase.
subordinator: "if"; "since"; "because"; "although".

#-------------------------SENTENCE PHRASE---------------------#
SVCO phrase: [circumstance],subject(NUMB,PERS),VCO phrase(NUMB,PERS).
SVOnC phrase(PREP): subject(NUMB,PERS),VOnC phrase(PREP,NUMB,PERS).
SVCnO phrase: subject(NUMB,PERS),VCnO phrase(NUMB,PERS).

#-------------------------NOUN PHRASE-------------------------#
subject(NUMB,PERS): noun phrase(NUMB,PERS,nom).
                    noun phrase(NUMB,PERS,CASE):
                    noun part(NUMB,PERS,CASE),rel phrase/"(",noun part,rel phrase,")";
                    noun part(NUMB,PERS,CASE)/"(",noun part,")".
noun phrase(plur,PERS,CASE): noun part(NUMB,PERS,CASE),coordinator,
             noun phrase(NUMB1,PERS1,CASE)/"(", noun part,coordinator,noun phrase, ")".
noun part(NUMB,third,CASE): [ART(NUMB)],noun group(NUMB,CASE),
                            (participle(TRAN,PREP,papl),CnO phrase(TRAN,PREP);
                                                                  circumstance; );
                            DET(NUMB),["of"],noun group(NUMB,CASE),
                            (participle(TRAN,PREP,papl),CnO phrase(TRAN,PREP);
                                                                  circumstance; );
                            noun group(sing,gen),noun group(NUMB,CASE),
                            (participle(TRAN,PREP,papl),CnO phrase(TRAN,PREP);
                                                                  circumstance; );
                            number(NUMB),noun group(NUMB,CASE),
                            (participle(TRAN,PREP,papl),CnO phrase(TRAN,PREP);
                                                                  circumstance; );
                            POSS PRON,noun group(NUMB,CASE),
                            (participle(TRAN,PREP,papl),CnO phrase(TRAN,PREP);
                                                                  circumstance; ).
                            noun part(sing,third,CASE):
                            "what",SVCnO phrase;
                            "that",SVCO phrase;
                            ["not"],participle(TRAN,PREP,prpl),CO phrase(TRAN,PREP);
                            "to",infinitive(TRAN,PREP),CO phrase(TRAN,PREP).
                            noun part(NUMB,PERS,CASE):
                            number(NUMB),"of ",noun part(NUMB,PERS,CASE);
                            PERS PRON(NUMB,PERS,CASE).
                            noun part(sing,third,CASE):
                            proper name.
proper name: name prefix, proper name;
             $MATCH("[A-Z][a-z]*"),[proper name].
name prefix: "Mr."; "Mrs."; "Miss."; "Ms.";
             $MATCH("[A-Z]\.").
noun group(NUMB,CASE): adjective phrase,noun group(NUMB,CASE);
                       participle(TRAN,PREP,prpl | papl),noun group(NUMB,CASE);
                       NOUN(NUMB,CASE);
                       NOUN(NUMB1,nom),NOUN(NUMB,CASE).
rel phrase: REL PRON(nom),VCO phrase(NUMB,PERS);
            REL PRON(gen),oSVC phrase;
            PREPOS(PREP),REL PRON(dat),SVOnC phrase(PREP);
            REL PRON(dat),SVOnC phrase(PREP),PREPOS(PREP);
            REL PRON(dat),SVOnC phrase(to);
            SVOnC phrase(PREP),PREPOS(PREP);
            [REL PRON(acc)],SVCnO phrase.
oSVC phrase: object without determiner,SVCnO phrase.
object compl(PREP): PREPOS(PREP),noun phrase(NUMB,PERS,dat).
indirect object(PREP): PREPOS(PREP),noun phrase(NUMB,PERS,dat).
indirect object(to): noun phrase(NUMB,PERS,dat).
modifier: PREPOS,(noun phrase(NUMB,PERS,dat); number(NUMB)).
object without determiner: noun group(NUMB,acc).
object: noun phrase(NUMB,PERS,acc).
object(trav|ditr): noun phrase(NUMB,PERS,acc).
```

```
predicate: adjective phrase;
           noun phrase(NUMB,PERS,nom);
           circumstance;
           participle(TRAN,PREP,papl),CO phrase(TRAN,PREP).

#-------------------------VERB PHRASE-------------------------#
VOnC phrase(PREP,NUMB,PERS): verb phrase(TRAN,PREP,TIME,NUMB,PERS),OnC phrase(TRAN);
                             to be(TIME,NUMB,PERS),[predicate],[circumstance].
VCnO phrase(NUMB,PERS): verb phrase(TRAX,PREP,TIME,NUMB,PERS), CnO phrase(TRAX,PREP);
                        to be(TIME,NUMB,PERS),[circumstance].
VCO phrase(NUMB,PERS): verb phrase(TRAN,PREP,TIME,NUMB,PERS),CO phrase(TRAN,PREP);
                       to be(TIME,NUMB,PERS),[predicate],[circumstance].
CO phrase(intr,PREP): [circumstance].
CO phrase(trav,PREP): object,[circumstance].
CO phrase(ditr,PREP): indirect object(PREP),object,[circumstance];
                      object,object compl(PREP),[circumstance].
CnO phrase(intr|trav,PREP): [circumstance].
CnO phrase(ditr,PREP): object compl(PREP),[circumstance].
verb phrase(TRAN,PREP,TIME,NUMB,PERS): [adverb],verb form(TRAN,PREP,TIME,NUMB,PERS);
             TO BE(TIME,NUMB,PERS),participle(TRAN,PREP,prpl);
             semiauxv(TIME,NUMB,PERS),"to",infinitive(TRAN,PREP);
             auxv(TIME,NUMB,PERS),infinitive(TRAN,PREP);
             TO HAVE(TIME,NUMB,PERS),participle(TRAN,PREP,papl);
             TO HAVE(TIME,NUMB,PERS),[adverb],TO BE(papl), participle(TRAN,PREP,prpl).
verb phrase(intr,PREP,TIME,NUMB,PERS): TO BE(TIME,NUMB,PERS),participle(trav,PREP,papl);
             TO HAVE(TIME,NUMB,PERS),[adverb],TO BE(papl), participle(trav,PREP,papl).
xSvCO phrase:
 TO BE(TIME,NUMB,PERS),subject(NUMB,PERS),participle(TRAN,PREP,prpl),CO phrase(TRAN,PREP);
 TO BE(TIME,NUMB,PERS),subject(NUMB,PERS),participle(trav,PREP,papl),CO phrase(intr,PREP);
 TO HAVE(TIME,NUMB,PERS),subject(NUMB,PERS),
       participle(TRAN,PREP,papl),CO phrase(TRAN,PREP);
 TO HAVE(TIME,NUMB,PERS),subject(NUMB,PERS),[adverb],
 TO BE(papl),participle(TRAN,PREP,prpl),CO phrase(TRAN,PREP);
 TO HAVE(TIME,NUMB,PERS),subject(NUMB,PERS),[adverb],
       TO BE(papl),participle(trav,PREP,papl),CO phrase(intr,PREP).
vSvCO phrase:
    auxv(TIME,NUMB,PERS),subject(NUMB,PERS),infinitive(TRAN,PREP), CO phrase(TRAN,PREP);
    to be(TIME,NUMB,PERS),subject(NUMB,PERS),[predicate], [circumstance].
vSvCnO phrase:
    auxv(TIME,NUMB,PERS),subject(NUMB,PERS), infinitive(TRAX,PREP),CnO phrase(TRAX,PREP);
    TO BE(TIME,NUMB,PERS),subject(NUMB,PERS),
       participle(TRAX,PREP,prpl),CnO phrase(TRAN,PREP);
    TO HAVE(TIME,NUMB,PERS),subject(NUMB,PERS),
       participle(TRAX,PREP,papl),CnO phrase(TRAN,PREP).
vSvOnC phrase(PREP):
    auxv(TIME,NUMB,PERS),subject(NUMB,PERS),infinitive(TRAN,PREP),
       [object(TRAN)],[PREPOS(PREP)],[circumstance];
    TO BE(TIME,NUMB,PERS),subject(NUMB,PERS),
       participle(TRAN,PREP,prpl),OnC phrase(TRAN);
    TO HAVE(TIME,NUMB,PERS),subject(NUMB,PERS),
       participle(TRAN,PREP,papl),OnC phrase(TRAN).
OnC phrase(TRAN): [object(TRAN)],[circumstance].

#-------------------------ADJECTIVE PHRASE--------------------#
adjective phrase: [adverb],ADJE(abso);
                  "so",adjective phrase,["that",SVCO phrase];
                  "as",ADJE(abso),"as",noun phrase(NUMB,PERS,CASE);
                  "more",ADJE(abso),["than",noun phrase(NUMB,PERS,CASE)];
                  ADJE(COMP),["than",noun phrase(NUMB,PERS,CASE)];
                  "most",ADJE(abso);
                  ADJE(supl).
adverb: ADVB.

#-------------------------NUMBERS-----------------------------#
number(sing): "one"; "1".
number(plur): "zero"; thousands; hundreds; tens, [digits]; twens; mores.
number (plur): $MATCH("[123456789][0123456789]\[1,\]");
               $MATCH("[23456789]").
thousands: [singles],"thousand",(hundreds; tens; singles; ).
hundreds: [singles],"hundred",(["and"],tens; ["and"],singles; ).
tens: "twenty"; "thirty"; "fourty"; "fifty"; "sixty"; "seventy"; "eighty"; "ninety".
```

```
singles: digits; twens.
twens: "ten"; "eleven"; "twelve"; "thirteen"; "fourteen";
       "fifteen"; "sixteen"; "seventeen"; "eighteen"; "nineteen".
digits: "one"; mores.
mores: "two"; "three"; "four"; "five"; "six"; "seven"; "eight"; "nine".

#-------------------------MORPHOLOGY-------------------------#
# or,rather,interface to the lexicon #
auxv(TIME,NUMB,PERS): AUXV(TIME),["not"].
verb form(TRAN,PREP,prst,NUMB,first | secnd): VERB I(TRAN,PREP).
verb form(TRAN,PREP,prst,sing,third):         VERB S(TRAN,PREP).
verb form(TRAN,PREP,prst,plur,third):         VERB I(TRAN,PREP).
verb form(TRAN,PREP,past,NUMB,PERS):          VERB PAST(TRAN,PREP).
verb form(TRAN,PREP,prpl): VERB PRPL(TRAN,PREP).
verb form(TRAN,PREP,papl): VERB PAPL(TRAN,PREP).
semiauxv(prst,NUMB,first | secnd): COP I.
semiauxv(prst,sing,third):         COP S.
semiauxv(prst,plur,third):         COP I.
semiauxv(past,NUMB,PERS):          COP P.
semiauxv(infi):                    COP I.
infinitive(TRAN,PREP): adverb,infinitive(TRAN,PREP).
infinitive(TRAN,PREP): VERB I(TRAN,PREP);
                       semiauxv(infi),"to",infinitive(TRAN,PREP);
                       TO BE(infi),participle(TRAN,PREP,prpl);
                       TO HAVE(infi),participle(TRAN,PREP,papl);
                       TO HAVE(infi),TO BE(papl),participle(TRAN,PREP,prpl).
infinitive(intr,PREP): TO BE(infi),participle(trav,PREP,papl);
                       TO HAVE(infi),TO BE(papl),participle(trav,PREP,papl).
participle(TRAN,PREP,TENS): [adverb],verb form(TRAN,PREP,TENS).
participle(TRAN,PREP,prpl): TO HAVE(prpl),verb form(TRAN,PREP,papl).
to be(TIME,NUMB,PERS): TO BE(TIME,NUMB,PERS),["not"];
                       semiauxv(TIME,NUMB,PERS),["not"],"to",TO BE(infi);
                       auxv(TIME,NUMB,PERS),["not"],[semiauxv(infi),"to"],(TO BE(infi);
                       TO HAVE(infi),TO BE(papl)).
```

A nyelvleírás elég tömör és nem hagy sok hiányt maga után. A rendszerben a megvalósíthatóság, hatékonyság és a tömörség volt a cél.

A többi a szótár dolga, amire néhány szótári bejegyzést mutatok:

```
"I"         PERS PRON(sing,first,nom)
"a"         ART(sing)
"after"     PREPOS
"always"    ADVB
"am"        TO BE(prst,sing,first)
"an"        ART(sing)
"any"       DET(sing)
"all"       DET(plur)
"appear"    COP I
"appeared"  COP P
"appears"   COP S
"are"       TO BE(prst,NUMB,secnd)
"are"       TO BE(prst,plur,first|third)
"arrow"     NOUN(sing,nom|dat|acc)
"arrow's"   NOUN(sing,gen)
"arrows"    NOUN(plur,nom|dat|acc)
"ask"       VERB I(ditr,from)
"asked"     VERB PAST(ditr,from)
"asked"     VERB PAPL(ditr,from)
"asking"    VERB PRPL(ditr,from)
"asks"      VERB S(ditr,from)
"at"        PREPOS(at)
"at"        PREPOS
"be"        TO BE(infi)
"been"      TO BE(papl)
"before"    PREPOS
"behind"    PREPOS
"being"     TO BE(prpl)
"best"      ADJE(supl)
"better"    ADJE(COMP)
"big"       ADJE(abso)
"bigger"    ADJE(COMP)
"biggest"   ADJE(supl)
"book"      NOUN(sing,nom|dat|acc)
"book's"    NOUN(sing,gen)
"books"     NOUN(plur,nom|dat|acc)
"boy"       NOUN(sing,nom|dat|acc)
"boy's"     NOUN(sing,gen)
"boys"      NOUN(plur,nom|dat|acc)
"brighten"  VERB I(intr,none)
"brightened"  VERB PAST(intr,none)
"brightened"  VERB PAPL(intr,none)
"brightening" VERB PRPL(intr,none)
"brightens"   VERB S(intr,none)
"bring"     VERB I(ditr,to)
"bringing"  VERB PRPL(ditr,to)
"brings"    VERB S(ditr,to)
"brought"   VERB PAST(ditr,to)
"brought"   VERB PAPL(ditr,to)
"by"        PREPOS
"came"      VERB PAST(intr,to)
"can"       AUXV(prst)
"clearly"   ADVB
"come"      VERB I(intr,to)
"come"      VERB PAPL(intr,to)
"comes"     VERB S(intr,to)
"coming"    VERB PRPL(intr,to)
"could"     AUXV(past)
"did"       AUXV(past)
"did"       VERB PAST(trav,none)
"didnt"     AUXV(past)
```

Az implementálás kihasználja, hogy véges raghalmaz részhalmazai is végesek, ezért a formalizmus teljes egészében átírható ragok nélküli környezetfüggetlen nyelvtanra. Sajnos a természetes nyelvek nem elemezhetőek egyértelműen, ezért remény sincs, hogy determinisztikus elemzővel gyors – az input hosszával arányos időigényű – elemzőt alkossunk. Általános elemzők között a Cooke-Younger-Kasami módszer és az Earley-módszer a leggyakrabban használatosak. Az előző használata a nyelvtan minimális átalakításával jár, bár ez nem igazán komoly plusz. A másik elvben bármilyen – felesleges szabályokat nem tartalmazó – nyelvtannál hatékony. Az AGFL-nél ez utóbbit alkalmazták. Az elvi megfontolások alapján mindkét elemző hatékonysága n hosszú input esetén $O(n^3*p*h)$, ahol p a nyelvtani szabályok száma, h pedig a szabályok jobb oldalának hossza. Elemzésnél a mondatok hosszával nem lehet mit kezdeni, de a szabályok számát lehetőleg alacsonyra kell tenni. Itt van jelentősége, hogy a raghalmazok nem minden részhalmazát kell kezelni, csak azokat, melyek a leírás alapján előfordulhatnak. Ha az elemzés során például az alany és az ige száma nem derül ki, (*you are*) akkor az elemző nem külön ágban, alternatívaként hozza ki a két elemzést, hanem közös szabályok alapján olyan lesz, melyben az alany és az ige elemzésében mindkét lehetőség szerepel.

Technikai, de általános trükk, hogy a nyelvtanból kiszűri a felesleges szimbólumokat, szabályokat, és a balrekurziókat, melyek némi gondot jelenthetnek az Earley-módszer második fázisában. Ezekre viszont bevált általános eszközök vannak.

http://www.cs.kun.nl/agfl/

8. Egyszerűsítő módszerek

Abban hiába reménykedünk, hogy tökéletes formális nyelvtan alkotunk. Már csak azért sem, mert a nyelv elég összetett ahhoz, hogysem matematikailag le lehessen írni. Azok az összetett leírások, melyek nagy mélységben foglalkoznak a nyelvi általános jelenségekkel, nem alkalmasak arra, hogy fel is töltsük tartalommal. Ha túl finom a nyelvi modell, akkor nem lesz, aki a szavakat, szabályokat feldíszítse a megfelelő kiegészítő információval. Gondoljunk arra, hogy egy jó szótárban rengeteg olyan információ van, melyek nem a szavak, hanem a mondatok fordítására kellenek: kifejezések, példamondatok, stiláris kiegészítések, vonzatleírások… Az ilyen információk egyes szavaknál hiányoznak, másoknál tévesek, és nem azért, mert ne tudná a szótárszerkesztő, hanem azért, mert egy ilyen szótár túllépi azt a méretet, mely karbantartható. Vagyis hiába javítják, bővítik, a százalékos hiba megmarad.

Ezek miatt nem feltétlen kell minden általunk ismert nyelvi jelenséget belevenni a nyelvtanunkba. Ha a célnak megfelel, ennél kevesebbel jobban elérhetjük célunkat. Mindig mérlegelni kell, az adott nyelvi tulajdonság kezelésének bevezetése mekkora munkával jár, mennyire pontosan lehet megvalósítani. Az angol nyelv elemzésénél a vonzatokkal kevésbé kell foglalkozni, ha a célunk csak a helyes/helytelen mondatok megkülönböztetése, míg a magyarban nagyobb súlyt kap a vonzatok kezelése.

A másik gond, hogy technikai és mentális egyszerűsítés miatt olyan szerkezeteket enged meg a nyelvi modell, amely a gyakorlatban soha nem fordul elő. Vagyis a modell kompetenciája nem mindig találkozik a performanciával. A leggyakoribb performanciai korlát a beágyazódás mélysége. A leírás – kompetencia – megengedi a tetszőlegesen mélyen egymásba ágyazódó almondatrendszert, birtokos szerkezetet, de a gyakorlatban a harmadik után már a szöveget befogadónak érthetetlenné válik a szerkezet. A nyelv formális leírása így viszont egyszerűbb ilyen egymásba ágyazások megengedésével.

8.1 Korlátozások a CF jellegű nyelvtanokban – véges automata, illetve véges implementációk (BUG)

Tipikus megközelítés, hogy egy reguláris nyelvtan technikailag jobban kezelhető, mint a környezetfüggetlen nyel. Ha igaz, hogy egy mondat hossza minden nyelvben korlátos – egy hossz után érthetetlenné válik – akkor minden nyelv mondatszintaxisa véges nyelv. Ebben az értelemben reguláris.

Akkor is erre a végkövetkeztetésre jutunk, ha az mondjuk, hogy a nyelvtani modell alapján meglevő rekurziók mélysége a performancia alapján korlátos – legyen a korlát 3 vagy akár 5 – a környezetfüggetlen nyelvtan ezzel a korláttal véges nyelvet ír le. Ezt azért ilyen erősen nem korlátoznám, hiszen az érthetőséget az egymásba ágyazódás mélysége korlátozza, de a nyelvtanban nem feltétlen egymásba ágyazódás okoz rekurziót. A lista is ilyen módon van megoldva a CF nyelvtanban. A felsorolásnál messze nem korlát az 5 elem. A lista szerkezete viszont feloldható reguláris nyelvtannal.

Bár kicsit más meggondolással erre a következtetésre jut Kálmán László és Prószéky Gábor, és erre építve készítették el a BUG (Budapest Unification Grammar) nyelvi leírást.

Lényeges, hogy a nyelv alapleírása környezetfüggetlen jellegű unifikációs nyelvtanra épít, melyre – mintegy rátét – bevezetik a mélységi korlátokat. Emiatt a leírás tömör marad, viszont lehetőség nyílik arra, hogy a nyelvtant átalakítsuk, pontosabban gépi módszerrel átírjuk reguláris jellegű nyelvtanra.

8.2 Lapos szintaxis véges automata alapon

A szintaktikus elemzés gyakori célja nem feltétlen az, hogy a teljes szerkezetet feltárjuk. Gyakran elegendő, ha a mondat alkotóelemeit felismerjük, elkülönítjük. Az AGFL-ben leírt mintaangol sem adja meg pontosan a részek közti viszonyt, de ennél lehet még egyszerűbb nyelvtant is használni. Tehát ha a célnak megfelel, akkor nem szégyen, ha gyengébb a nyelvtanunk, mint ami egyáltalán létre tudnánk hozni. Egyrészt csökkenthetjük a többértelműséget, másrészt akár reguláris nyelvtannal is megoldhatjuk a leírást. Angol nyelvre léteznek ilyen hasznos mondatelemzők, de magyarra is elképzelhető olyan reguláris nyelvtan, amely a névszói kifejezéseket, az igei szerkezeteket jól elkülöníti, de azok közti összefüggést nem tárja fel. Az **5.1**-ben tárgyalt névszói modell kis fáradtsággal reguláris formára írható át. Hasonlóan az összetartozó igei kifejezés szintaxisa is. Ha a távoli egyeztetéseket nem vesszük figyelembe, korlátozzuk az egymásba ágyazódást, akkor az egész mondatszerkezet leírható reguláris nyelvtannal.

Akkor is ilyen módszerhez folyamodunk, ha nincs, megfelelő mondattani leírásunk, vagy előállítása drága. Például célzott tartalomkivonatoláshoz elegendő a mondat néhány kulcskifejezését megtalálni, és ehhez untig elég felismerni, elkülöníteni a mondat fő részeit.

8.3 Lokális, illetve parciális szintaxis haszna egyértelműsítéseknél, szövegellenőrzésnél

Van, amikor egyenesen nincs szükség az egész mondat elemzésére. Ennek oka lehet az, hogy nincs megfelelő minőségű nyelvi leírás arra, hogy időben az egész mondatot elemezzük – az elemzés ideje a szabályok számával és az elemzendő sztring hosszának köbével arányos. Ha a célnak megfelel a mondat egy részének elemzése, akkor kisebb egységeket, például igei, névszói kifejezéseket lehet magában elemezni. Gyakran ennél kisebb egységet is részként lehet elemezni. Ezeket lokális szintaxisnak nevezzük.

Magyarban – a kifejezések szétszakadás miatt – olyan eszközre is szükség lehet, melyben nem teljes elemzéssel felismerünk olyan összefüggéseket, melyek lokalizálhatók a mondat folytonos szakaszában. Ezek a nyelvtanok nem a mondat egészét írják le, csupán a nyelvi jelenségek egy-egy szűkös jelenségét. Az ilyen nyelvtanok a parciális nyelvtanok.

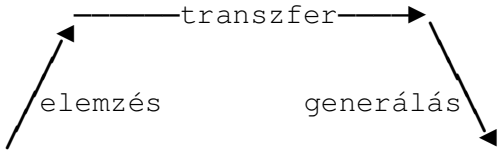
Hasonló a gond és a megoldás, ha például a mondatban egy, a rendszer által ismeretlen vagy helytelenül írt szó szerepel, esetleg nyelvileg helytelen mondattal kell valamit – minél többet – kezdeni. Ilyenkor egy teljes elemző széttárja a kezét, és azt üzeni, nem értelmezhető a mondat. Pedig élő beszédben is felfogjuk azokat a szövegeket, melyben általunk sohase hallott szavak szerepelnek.

8.4 A szintaxis szerepe a gépi fordításnál

A számítógépes nyelvészet elsődleges célja a gépi fordítás megvalósítása volt. A mintegy hatvan éves történetében sok minden megvalósult, de a kezdeti idea még csak az utóbbi évtizedben válik valóssá.

Amikor az ötvenes években megjelentek a mondatok szerkezetét feltáró programok, a fő gondolat a következő volt. A gépi fordítás három fázisból áll:

1. A forrásnyelven elvégzik a mondat szintaktikus elemzését, melynek eredménye egy levezetési fa.
2. A célnyelv szintaxisa eltérhet ettől, ezért egy második lépésben ezt a szintaktikus szerkezetet kell átírni, transzformálni a célnyelv szintaktikus szabályainak megfelelő szerkezetre.
2. Ha ez megvan, akkor ebből a szerkezetből és a szavak egyedi fordításából kigenerálható a célnyelvi szöveg.



forrásszöveg célszöveg

Az elképzelés jónak tűnt, és sokáig tartotta magát az elemzés-transzfer-generálás irányzat. A kezdeti sikertelenséget az akkori számítástechnika szűk voltának tudták be. Még a nyolcvanas években is sok ilyen kísérlet volt. A szintaktikus elemzők kifinomultak, a nyelvtani modellek bonyolódtak, de ezen az alapon használható módszer nem született. A 80-as 90-es években inkább az volt a kérdés, milyen magas szintű legyen az elemzés, vagyis a mondatok absztrakciója milyen mély legyen. Ha kellően absztrakt, akár nyelvfüggetlen, akkor a transzfer fázisára kevés feladat jut. Ha alacsony, akkor viszont könnyebb megfogalmazni az elemzés és a generálás szabályait. A magas absztrakciós szintet főként európai projektekben tűzték ki célul. Hiszen, ha kellően mély elemzést adunk, akkor még az is elképzelhető, hogy bármely európai nyelvpár esetén független a mondatok elemzése és a generálása a másik nyelvtől, hiszen a mondatok elemzés után kapott reprezentációja nyelvfüggetlen, illetve nyelvfüggetlen reprezentációból kell generálni a célnyelvi szöveget.

A tapasztalat nem támasztotta alá ezeket az elképzeléseket. Igaz az, hogy az absztrakt reprezentációt nem sikerült a nyelvtől teljesen függetlené tenni. Lehet, hogy ennek az az oka, hogy az ember által érthető jelentés csak a nyelv által létezik, emiatt nem is lehet független a nyelvtől. A másik lehetséges ok, hogy a nyelv nem csak jelentést hordoz. Rengeteg olyan tulajdonsággal rendelkezik beszédünk, melyek inkább kulturális, szokás, hagyomány miatt olyan, amilyen. Ezt viszont egyszerű formalizmussal nem lehet feltérképezni.

Ennek ellenére a szintaktikai elemzésnek nagy szerepe van a fordításban. Szerencsére – főként a 60-as években – kialakult egy olyan elméleti módszer, amely a fordítást elősegíti. Ez a fordítást jellemző nyelvtan fogalma köré épül.

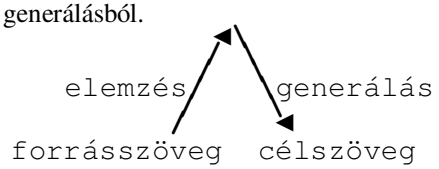
Már az **4.12.5** fejezetben is szó volt a fordítást jellemző nyelvről. A környezetfüggetlen nyelvek fordításához használják a fordítást jellemző nyelvtant. A programnyelveknél gyakran használatos módszer lényege, hogy fordításnál a forrás- és a célnyelvtan nem független, hanem szinkronban vannak. Pontosabban a két nyelvtan grammatikai jelei (természetes nyelvnél a nyelvtani fogalmai) azonosak, és levezetési szabályai is úgy vannak párba szedve, hogy a szabályok bal oldala (tehát a kibontandó nyelvi fogalom) azonos mindkét szabályban, és a levezetési szabály jobb oldaliban is egy-egy értelműen meg vannak feleltetve a forrás- és célnyelv szabályának grammatikai jelei. Tehát nem csak ugyanolyanok és ugyanannyian vannak a szabály grammatikai jelei, hanem azt is jelölik, melyik melyiknek felel meg. Ez utóbbira akkor van szükség, ha a jobb oldalon egy grammatikai jelből több is szerepel.

Programnyelvi egyszerű példa az aritmetika infix és postfix jelölésének átírására készített nyelvtanpár:

E→E+E	E→EE+
E→E−E	E→EE−
E→T	E→T
T→T*T	T→TT*
T→(E)	T→E
T→a	T→a

Most nem jelöltem, hogy a levezetés jobb oldalán melyik E melyik E-nek felel meg, hanem feltételezem, hogy az első az elsőnek, a második a másodoknak. Ha megkapom a forrásnyelv, az infix aritmetikai kifejezés egy elemzését, akkor ennek alapján a célnyelv levezetési fáját is megkaphatom könnyen, ha követem, melyik forráslevezetés szimbólumán milyen szabályt alkalmazok. A terminálisok akár mások is lehetnek. Sőt, a postfix aritmetikában nincs zárójel, míg az infixben van. A megfelelő kifejezések levezetési fája egyébként azonos szerkezetű. Tehát, ha megvan az egyik nyelvben a levezetési fa, akkor a másikban is, ezért a fordítás fő attrakciója az elemzés. A transzfer triviális.

Az adott példában az is jó, hogy a szinkronizált szabályokban a megfelelő grammatikai jelek megtartják sorrendjüket a szabályokban. Ha ez igaz a szintaxisvezérelt fordítási sémában, akkor azt mondjuk, hogy ez a séma a fordítást szigorúan jellemzi. Ebben az esetben nincs szükség transzferre, tehát a fordítás két fázisban történik: elemzésből és generálásból.



Igaz az a tétel, hogy két nyelv között akkor és csak akkor létezik veremfordító, ha létezik a fordítást szigorúan jellemző nyelvtan. A természetes nyelveknél ez nem várható. A szigorú séma azt jelentené, hogy ha a szavak nem is azonosak, de a mondat konfigurációja, tehát a szósorrend megegyezik a két nyelvben.

Ha két hasonló nyelv között kell fordítani – például latin nyelvek között: spanyol, portugál, olasz, francia, vagy szláv nyelvek közt: horvát, szerb, szlovák, cseh – akkor van remény arra, hogy szintaxist szigorúan jellemző fordítási sémát alkossunk, mely alapján egyszerű a fordítás. Vegyük észre, hogy ebben az esetben nincs szükség a mondat bármiféle mély, esetleg jelentéstani elemzésére. Ha a két nyelv azonos szerkezettel ír le sok-sok jelenséget, akkor ezek szerkezetét nem kell megkülönböztetni. Az ebben szereplő forrásnyelvi nyelvtan egyszerűbb, mint az absztrakt szerkezetet is leíró nyelvtan, de függhet a célnyelvtől. A célnyelv nyelvtana viszont nem kell, hogy teljes legyen, lehet szűkebb. Mint az emberek által használt nyelvnél, itt is igaz, hogy a generált (aktív) nyelv szűkebb, egyszerűbb, mint az elemzendő (passzív) **4.3**.

Általában két nyelv között, még ha létezik is fordítást jellemző fordítási séma, nem valószínű, hogy ez szigorúan jellemző. Ha magyarról angolra fordítunk, biztos, hogy nincs, hiszen semmi sem kötelezi a magyar beszélőt, hogy az alany megelőzze az igét, míg az angolban ez kötelező. Ennek ellenére igaz, hogy ha sikerül megelemezni az egyik nyelven a mondatot, akkor könnyen generálható a séma alapján a fordított mondat.

Érdekes módon ezt a módszert nem nagyon alkalmazzák az élő nyelveknél. A szintaxisvezérelt fordítási sémának az egyedüli sikeres használata az angol-magyar és magyar-angol irányba használt MetaMorpho. A generatív modellen alapuló fordítások inkább azt a módszert használják, amelyben az elemzés és a generálás közt transzfert alkalmaznak. De nézzünk példát a szintaxisvezérelt fordítási sémára:

```
*VP=suggest+NP+PRED:7
EN.VP= TV(lex="suggest", :passtr=ACT) + NP + NFSP(tensed=BAREINF)
HU.VP= N[lex="az", case=ACC] + TV[:lex="javasol"] + NP[case=DAT] + PUNCT[lex="comma"] +
       CONJ[lex="hogy"] + NFSP[:ptense=FELSZ]
```

A fenti MetaMorpho szabálypárban a `VP, TV, NP NFSP` grammatikai jelek vannak szinkronban. A csak magyar oldalon szereplő `N, PUNCT és CONJ` tulajdonképpen terminálisok, vagyis a felszíni alakjuk egyértelműen kigenerálható. Így megfelel a fordítást jellemző nyelvtan kritériumának.

A szabály alapján annak a mondatnak a fordítása, hogy *I suggest Peter phone him.*, *Azt javasolom Péternek, hogy telefonáljon neki*.

```
*VP=do+to+NP:10
EN.VP= TV(lex="do", :passtr=ACT) + ADJ + NP[lex="thing", case=NOM] + PREP[:lex="to"] + NP
HU.VP= ADJ[case=NOM] + TV[:lex="tesz"] + NP[case=INSTR]
```

Itt az angol szabályban vannak terminálisba átmenő többlet (`PREP`), míg a többi szinkronban van. Pl. *You have done bad thing to the Earth.* fordítása *Rosszat tettél a Földdel*.

9. Statisztikai módszerek a mondattannál

A generatív módszerre épülő nyelvtani alapú módszerek, ha egy kicsit is kifinomultak, azt eredményezik, hogy a mondatoknak általában mérhetetlenül nagyszámú elemzését kapjuk. Olyankor is, amikor egy ember – hallva a mondatot – egyértelműen értelmezi azokat, a gép meglepő alternatívákat talál. Ez gondot okoz az elemzésben, és legyen bármilyen kifinomult, cseles a fordító, még a fordítás időigénye is megnőhet annyira, hogy már zavaró legyen. A programnyelvek fordításánál ilyen gond nem merül fel, mert azok szintaxisát igyekeznek egyértelműre definiálni, sőt igény az is, hogy determinisztikus elemzővel lehessen fordítani. A gépi nyelvekre emiatt létezik lineáris sebességű fordító. (A bemenet hosszával arányos időt vesz igénybe az elemzés.) A természetes nyelveknél ez lehetetlen. Az elemzőkben ez ellen a következő módszerekkel védekezhetnek:

1. Elnyomás: rövidre vágással egy sikeres részelemzés letilthat más alternatívát. Ezt alkalmazza például a MetaMorpho, de hatékony Prolog algoritmusokban is ezen az alapon egyértelműsítenek.
2. Valószínűségeket társít a szabályok alkalmazásához, így a lokális valószínűséggel globális optimum felé befolyásolja az elemzést. Ezt alkalmazza az AGFL.
3. Az elemzés döntéseit pozitív, negatív súlyokkal pontozza, így a végső elemzési alternatívákat súlyozza. Ezt alkalmazza az **5.9-**ben említett kísérleti mondatelemző.

A korábbi módszerek arról szóltak, hogy egy szövegről igen-nem választ kapunk, vajon megfelelő-e. Akkor is, ha valószínűségi alapon döntünk, a végeredmény azon kívül, hogy kapunk egy elemzést vagy egy fordítást, valamilyen mértékszámot is rendel hozzá az algoritmus. Egyes felfogások szerint az, hogy egy szöveg helyes-e, illetve mi lehet az elemzése, az valószínűségi kérdés. Egy helytelen, de teljesen érthető mondatot magyarnak tekinthetünk, de kevesebbre értékeljük magyarságát, mint a mondat helyes változatának. Ha azokat a szövegeket akarjuk kezelni, melyek a valóságban előfordulnak, akkor ettől nem szabad ide-

genkedni. Már csak azért sem, mert a nyelvtanunk sem szentírás. Az is csak egy modellje a nyelvnek, és nem azonos vele.

A mondatokat kompetens, nyelvhez értő emberek kiértékelhetik, de nem az összes magyar mondatot. Emiatt a válasz arra a kérdésre, hogy mekkora valószínűséget rendeljünk az egyes mondatokhoz, lehetetlen jó választ adni. A lehetséges mondatok száma végtelen. Már a szavak kiértékelése is nehéz feladat, lásd a szavak gyakoriságáról szóló Zipf törvényt. A mondatok száma oly sok, hogy gyakrabban hallunk olyan mondatot, amelyet még sohasem, mint olyat, amit valaha már láttunk, hallottunk. Emiatt az az óhaj, hogy közvetlenül a mondatok valószínűségével kezdhetünk valamit, hiú remény. A Zipf törvényhez hasonló viszont a mondatoknál is létezik, csak az a baj, hogy míg a leggyakoribb 100 szó lefedi a szövegek felét, a leggyakoribb 10 000 mondat a szövegek egy százalékát sem teszi ki.

Hogy ezt áthidaljuk, két lehetőség van:

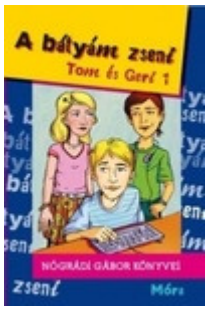
1. Csökkentjük a nyelv terét, amelyben valószínűségi alapon akarunk dönteni valamiről.
2. Nem mondatok, hanem kisebb egységek valószínűségével foglalkozunk.

Mindkét esetre létezik teljesen végletes megoldás. Elfeledkezünk a nyelvtanról, csupán a tényleges előfordulások döntenek. Tehát egy elvi kompetencia alkalmazása helyett, melyet úgy sem biztos, hogy pontosan tudunk formalizálni, csak azt nézzük, mi az, ami valóban előfordult már, és ennek ismeretét terjesztjük ki a további döntéseinkre.

Az első esetet alkalmazzák, ha nem a teljes nyelvet, hanem egy szűk részterületét választjuk. Már a szavak szintjén is kimutatható, hogy egy jó tollú író a nyelv szavainak legfeljebb egytizedét használja. A mondatok terén ez mégúgy igaz, hát, még ha egy cég műszaki leírásainak kötött terminológiát használó nyelvét nézzük.

A második esetben nem törődünk a mondatokkal, hanem például egymás utáni szópárok, szóhármasok… előfordulásának gyakoriságait vetjük össze. Ezt utóbbi esetben az gondolhatnánk, hogy a statisztikai nyelvi modell tere is nagy, hisz ha csak 10 000 szó lenne, akkor is 10^{12} a különböző hármasok száma.

9.1 Memória alapú fordítások – fordítómemóriák



Nemrég jelent meg egy aranyos gyermekregény Nógrádi György tollából, *A bátyám zseni* címmel, aminek az az alaptörténete, hogy egy gyerek elkészíti az angol-magyar gépi fordítót. Nem kell mást tenni, csupán leírni az összes angol mondatot, és melléírni a fordítását. A regényíró persze nem számolt azzal, hány mondatot kell leírni. Ha megbecsüljük, hogy a legfeljebb ötszavas angol mondatokból is 1000 000 000 000 000-nél több van, akkor a kedves főhősnek a matuzsálemi kor is kevés, mindet bepötyögni a gépbe. Bár, a regényben a kedves nagymama is besegít. Nem igazán kivitelezhető megoldás. A regénynek nem kell realisztikusnak lennie, de a fordítás feladatát mégis segítheti egy ehhez hasonló megoldás. Tudniillik, ha a mondatokat a valószínűségük sorrendjében rögzítjük, akkor a gyakoribb mondatok már feltehetően benne lesznek a gyűjteményben. Legalábbis már mérhető hányadát megtalálhatjuk, főként, ha nem a teljes nyelvet akarjuk lefedni, hanem egy szűk részét. Ezen az alapon jöttek létre azok a segédletek, amelyek az emberi fordítást segítik Talán először a *Kereskedelmi levelezés* tárait készítették el, melyek a gyakran alkalmazott levelezési mintákat adták meg különböző nyelveken. A felhasználónak csak ki kellett keresnie például azt a formát, amivel a késedelmes szállítás miatt elnézést kérnek a kereskedelmi partnertől.

Ezeket a mintákat emberek gyűjtötték össze. A fordítómemória lényege, hogy nem kell erőlködni a gyűjtéssel. A gép beépülve a fordítási rendszerbe gyűjti az emberi fordításokat, és ha előkerül egy olyan mondat, melyet már valaki valamikor a rendszerben lefordított, akkor annak a korábbi fordítását tálalja.

A kereskedelmi levelezésnek nagyon sablonos a nyelvezete, de azért ott sem azonos két levél. Ha más nem, a partner neve, az ügyintéző, az áru, az ár valamelyike biztos különbözik. Ezeket a fordítási mintában könnyű emberi erővel pótolni. A fordítómemória esetén is lehet, hogy egy mondat csak majdnem olyan, mint a korábbi. Ezek száma sokkal nagyobb, mint azoké, melyek szóról szóra megegyeznek egy korábban előfordulttal. Márpedig ha egy gépi fordító az esetek felében sem segít, akkor jobb nem használni. Ha viszont azonosítani tudjuk, hogy a forrásnyelv eltérő része a célnyelv mondatának mely részének felel meg a korábbi fordításoknál, akkor könnyen lehet, hogy egy-két szó lefordításával kipótolhatjuk a hiányt.

Tehát a fordítómemória két adatbázisra épül. Az egyik a lefordított mondatpárok tára, a másik a kisebb egységek fordításának tára, melyet leggyakrabban terminológiai szótárnak hívnak. Ennek alapján a fordítás

menete az, hogy a bejövő szövegre megkeresi a rendszer a leginkább hasonlító forrásmondato(ka)t, és ha nem teljes az egyezés, akkor a mintában eltérő részt kikeresi a terminológiai tárból, és azt a fordításban kipótolja. A humán fordító az így kapott nyersfordítást ellenőrzi, és ha megfelel, helyben hagyja, ha nem, akkor kijavítja. Az új fordítás bekerül a fordítások tárába, az új behelyettesítés pedig növeli a terminológiatárat.

A módszer azoknál az alkalmazásoknál eredményes, ahol valóban szűk a nyelv, és sokat, gyorsan kell fordítani. Ilyenek például a terméket gyakran fejlesztő cégek igénye, melyeknél az új termék leírását, karbantartási utasításait évről évre átírják, de sok nyelven kell megjelentetni. Például repülőgépek karbantartási leírása, szoftverek felhasználói kézikönyve…

A világon sok ilyen program létezik, hisz algoritmusa nyelvfüggetlen, teljesen matematikai alapon megfogalmazható. A legelterjedtebb a Trados, de létezik ennél jobb is. Magyarországi fejlesztés például a MemoQ. Ma már nem is annyira maga a fordító algoritmus, amiben jobb vagy kevésbé jó egy ilyen rendszer, hanem az azt körülvevő környezet, szolgáltatás. Mennyiben terheli le a felhasználót az a plusz, amivel növekszik az adatbázis, milyen lehetőség van különböző forrású terminológiai tárak importjára, exportjára, lehet-e a fordítási rendszert beépíteni más rendszerekbe, lehet-e automatikus adatfeltöltést végezni párhuzamos (többnyelvű) korpuszokból, lehet-e ellenőrizni a fordítás minőségét, konzisztenciáját, hogyan lehet ellenőrizni a fordítási folyamatot, karban tartani, felügyelni a gyűjtött adatbázisokat…

9.2 Statisztikai fordítók

Míg az előző módszer alapvetően szűkített nyelv esetén hatékony, a statisztikai fordítás teljesen általános céllal is lehet használni. A gond ott van, hogy az alapegységek, a mondatok tere annyira nagy, hogy a statisztika (az előfordulási gyakoriság mérésének) megbízhatóságával semmire sem megyünk, mivel a mondatok többsége jó, ha egyáltalán előfordul korábbi szövegekbe, de az előfordultak száma is általában egy. Ebből használható statisztikát nem lehet kapni. Az ötlet hasonlít ahhoz, mint amit a karakterstatisztikánál használtunk. Ott sem kellett teljes szavak gyakoriságát mérni, elég volt betűkettesek, -hármasok valószínűségét becsülni. A szósorozatok tere, melyben statisztikát készítjük, elvileg nagy. Már szóhármasokból is milliárdokat lehet összeállítani egy nyelven, de ezek nagyon kis hányada – ez is több millió – ami valóban előfordul egy nyelv összes fellehető korpuszában. A statisztikai felmérés abból áll, hogy párba kell állítani az egyik nyelv szavait, szópárait, szóhármasait a másik nyelv előforduló kisebb szósorozataival, és ebből kell statisztikát készíteni. Ha magyar szövegben mindig megjelenik a *házi feladat* ha az angolban a *home work*, németben a *Hausaufgabe*, oroszban a *домашнее задание*, akkor egyrészt ezen szócsoportok nagy gyakorisága miatt megfelelnek egymásnak, akkor is, ha németül csak egy szó szerepel. A *feladat* szónak a *task* megfelelője lehet, hogy nagyobb gyakorisággal szerepel a korpuszokban, mint a *work*, de ebben a közvetlen környezetben nem.

A statisztikai fordítók nyelvi modellje abból áll, hogy a forrás és a célnyelvi szócsoportok közti valószínűséget térképezi fel már meglevő anyagból, és ennek alapján – a szokásos naiv Bayes módszerrel – próbálja egy új forrásnyelvi mondathoz megtalálni a legnagyobb valószínűséggel illeszkedő célnyelvi szöveget.

A statisztika lehet kötött szórendű – ebben az esetben csak az egymást szorosan követő n darab szó lehet a statisztika tárgya, és lehet szabad szórendű – ebben az esetben a szinkronmondatok minden fellelhető, nem feltétlen egymás mellett álló szócsoportjaira készül el a statisztika.

A statisztika alapján lehet megkísérelni egy új mondat fordítását. Az eljárás az, hogy a célnyelven olyan szósort kell létrehozni, amelyben a forrásnyelvi szöveg háromszavas részsorozataival összevetve a valószínűségek összege vagy szorzata a legnagyobb, vagy ha szabad szórendű statisztikát használunk, akkor arra épít a valószínűségi kiértékelés. Ez utóbbi azoknál a nyelvpároknál fontos, ahol a mondat szerkezete lényegesen eltér. Ilyenkor a szavak sorrendjét más módszerrel kell meghatározni. Persze itt is lehet statisztikai módszert alkalmazni. Egynyelvű korpuszból a szavak sorrendjét lehet kiértékelni, ahogyan a betűstatisztika alapján fiktív szavakat. Vagyis a kapott célnyelvi szósorozat közül azt kell választani, amelyik a legnagyobb valószínűséget kapja a részhármasok kiértékelésének összesítése alapján. Ez általában nem elég. Lásd **9.4**.

A statisztikai módszerek alapja is a nagy minta megléte, akár a fordítómemória esetén. Elvileg, ha a minták hossza nő, valamint a minták mennyisége minden határon túl nőhet(ne), konvergálnánk a tökéletes fordítás felé. Itt ugyanazokba az elvi korlátokba ütközünk, mint amelyek már a szótannál is jelentkeztek. Ha a vizsgált tér elég nagy, akkor a gyakori esetek valószínűségét jól tudjuk becsülni az előfordulásuk gyakorisága alapján. De már a szóhármasok is akkora teret alkotnak, hogy a ténylegesen előforduló esetek csak a töredéke a lehetséges, akár még pozitív valószínűséggel is felruházható eseteknek. Emiatt a statisztika

megbízhatósága ugyancsak hagy maga után kívánnivalót. Arról nem is beszélek, hogy nagy korpusz esetén a hibás anyag – értsd félrefordítások – százaléka nem hogy javítaná, hanem rontja a minőséget. Márpedig ha nagy korpuszt akarunk, akkor ellenőrizetlenül kell bevenni a fordítások tömkelegét.

A statisztikai nyelvi modell felépítése teljesen általános matematikai eszközt igényel. Emiatt előszeretettel és sikerrel alkalmazzák. Szabadon elérhető a Moses, de sokan ismerik a Google fordítószolgáltatását, vagy a Microsoft Bing projectjét. Ez utóbbi kettő sok nyelvpárra nyújt eszközt, és kellően nagy adatbázis van mögöttük, hogy hasznos legyen. Egyértelmű, hogy korpuszok méretében a Google vezet. Becslésem szerint ötször akkora anyagot dolgoztak fel, mint bárki más. A fordítás minősége mégsem annyival jobb, mint például a Bingé. Azért nem, mert elérték azt az elvi korlátot, amikor már hiába van tízszer, százszor akkor mintájuk, a fordítás minősége nem válik mérhetően jobbá.

Amikor megjelentek, akkor egyből kiderült a statisztikai fordítás előnye és hátránya is:

1. **Előnye**, hogy a performanciát a legvégsőkig kihasználja. A *home work* véletlenül sem lesz *otthon munka*, és a fordításban olyan megoldásokat talál, amelyek nagy valószínűséggel előfordulnak a másik nyelven is. Valóban a generatív módszerek hátránya, hogy a sok lehetőség közt nem tud igazán választani. A nyelvnek olyan sok jelensége van, hogy azokat maradék nélkül nem lehet számba venni, formalizálni. Nem csak nyelvtan létezik, létezik stiláris ismérv, nyelvszokás, kulturális különbség, néha még a szó hangzása is közrejátszik, miért így, miért nem úgy fordítunk. Ha a német azt mondja, hogy *Malzeit*, akkor mi *jó étvágyat kívánunk*. A buszon az a magyar szöveg, hogy *Ne támaszkodjon az ajtónak!*, angolul csak annyi olvasható, *mind the opening door*, oroszul pedig *Не прислоняться к дверям!*, ami szó szerinti fordításban *gondoljon a nyíló ajtóra*, illetve *ne elefántkodjék az ajtónak*, stilárisan *ne tehénkedjen az ajtónak!*. lenne. A *nagyon sok* kifejezése is eltér nyelvenként. A magyar kis nép, az orosz nagyobb létszámú, a kínaiak még többen vannak. Ezért ha magyarul azt mondjuk, *százával jön az ellen*, akkor oroszul *ezrével*, kínaiban meg *millió* a nagyon sok. Ami nekünk *Közel-Kelet*, az Nyugat-Európának *Közép-Kelet*. Nos, az ehhez hasonlókat simán kezelheti a statisztika.
   A hasonló nyelvek közt remek fordításokat lehet elérni. Francia olasz, spanyol közt egész elfogadható a fordítás minősége. Itt a nyelvtan kevésbé számít, hisz a szósorrend, a nyelvtani hasonlóság miatt ezekre általában nem kell külön gondot fordítani. A ténylegesen gyakoribb mondatok esetén meglepően jó fordítást kapunk, még akkor is, ha a két nyelv lényegesen máshogy formálja ugyanazt a mondatot.
2. **Hátránya**, hogy a statisztika félrevihet. Az Google fordító első verziójában akár melyik nyelvről fordítottunk angolra, vagy vissza, az *English* szónak mindig az volt a megfelelője amilyen nyelvre fordítottak. Franciáknál *français*, olaszoknál *italiano*, magyaroknál *magyar*. Hát persze! A legkénnyelmesebben hozzáférhető kétnyelvű korpusz, a kereskedelmen kapható termékek soknyelvű leírása, ahol az első szó a megfelelő nyelvű rész nyelvének megnevezése volt. Ami az angol nyelvű részben *English*, a magyarban *magyar*. Bizonyos szituációkban ez jó is. Ha azt fordítom más nyelvre, hogy *magyarán szólva*, nyílván – *auf gut deutsch*, *in plain English*, *говоря по-русски,* és nem a magyart kell emlegetni. Általában viszont meg kell hagyni a nép nevét. Ezt a hibát ugyan gyorsan korrigálták, de hasonló hiba ezrével akadhat.
   A statisztika máshogy is félre visz. Ha egy gyakori mondaton egy kicsit változtatunk, ahogy még sohasem írták le, akkor szinte biztos, hogy nem a fordítandó mondat jelentését kapjuk vissza, hanem gyakori mondatot. Feltételezem, hogy annak a mondatnak, hogy *kutyából lesz szalonna* a Google azt az értelmet fordítja, ami a *kutyából nem lesz szalonna* jelentést bírja.
   A leggyengébb pont a nyelvtan hiánya. Emiatt lényegesen eltérő mondatoknál a látszólagos jó részfordításokból rossz egész áll össze.
   További nehézség azoknál a nyelveknél fordul elő, ahol a szóalakok száma nagyon nagy. Mivel az algoritmus tisztán szóalakokra épül, semmi morfológia nincs benne, a szóhármasok tere nagyságrendekkel nagyobb, mint az angol, vagy akár a latin nyelveknél. Ez viszont – hasonló megbízhatóságú statisztikát keresvén – több nagyságrenddel bővebb korpuszt kíván a feldolgozáshoz. Az biztos, hogy magyar-angol, török-angol nyelvű korpusz soha nem lesz akkora sem, mint például francia-olasz, vagy holland-angol, de nem ez az igazi korlát, hanem az, hogy a korábbi nyelvek szószerkezete összetettebb.

9.3 Shake and Bake

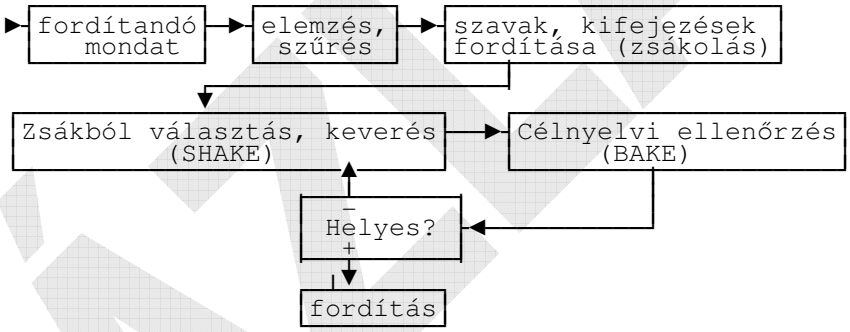
A nyelvfüggetlen struktúra, szemantika, mint interlingua reprezentáció a fordításnál nem vált be. A fent említett szintaxis vezérelt fordítási sémának egyik nagy hátránya, hogy a leírt nyelvtan a gyakorlatban nyelvpárfüggő, emiatt meglehetősen nagyméretű, nehezen karbantartható. Mindkettőnél az a nagy gond, hogy a nyelvnek csak egy része a nyelvtan és a szemantika. A nyelvtan használható formalizálására van remény, a szemantika már keményebb dió. A gond ott van, hogy a szemantikán és szintaktikán kívül sok más is befolyásolja a fordítást. Ezek feltérképezése, absztrakt leírása lehetetlen. Emiatt igaz, hogy csak azok a fordítók hatékonyak, melyek ráépülnek a két nyelv egyedi tulajdonságaira.

Ez azzal jár, hogy külön munka marad nyelvpáronként a transzfer, de legtöbb esetben a forrás és célnyelv nyelvtani leírását is befolyásolja a másik nyelv. Ebből próbál kitörni az a módszer, amellyel a 90-e évek óta próbálkoznak.

A módszer alapgondolata az, hogy a forrásnyelv és célnyelv szintaxisát függetlenül írják le. A transzfer eszköze lehet akár statisztikai fordító is, de lehet egy kétnyelvű szótár alapján működő durva fordító is. Mindkét esetben az elemzett mondat szavaira, esetleg kifejezéseire nem egy fordítás létezik, hanem egységenként egy zsák megoldást lehet vagy statisztikai alapon, vagy szótár segítségével összegyűjteni.

Az így kapott részfordítások nem feltétlen egyeznek meg a célnyelvi mondat sorrendjének, ezért az így kapott egységeket még meg lehet keverni, majd a kapott eredményt kell alávetni a célnyelvi elemzőnek. Ha a kapott eredmény megfelel a célnyelvi leírásnak – megtartva a kívánt egységek részfordításból származó tulajdonságainak –, akkor várhatóan egy jó fordítást kapunk.

Tehát a négy függetlenül használható modul. a forrás és célnyelvi elemző, a fordítólexikon és a keverő:



Ha feltételezzük, hogy az elemzők kellően gyorsak, akkor a módszer sebessége attól függ, mennyire lehet csökkenteni a részfordítások számát, és a sorrendiséget. Ha a lexikon által kiválasztott részfordításokat csak akkor tesszük a zsákba, ha azok a forrásnyelvi elemzésnek is megfelelnek, valamint a nagyobb megtalált egységeket egybe hagyjuk a sorrendi keverésnél, akkor némileg gyorsul és pontosodik a fordítás. Ha a zsákból való kiválasztást valószínűségek alapján sorrendezzük, az is sokat segít.

A shake and bake-re épülő deszkamodellek, kísérleti fordítók azt mutatják, hogy nem elveszett irányzat. A fő előnye, hogy mind a forrás, mind a célnyelv nyelvtanára elég egy robusztus formalizmust megadni. A transzfer alapja lexikon jellegű információ, sok kész szótárból kinyerhető. Az algoritmus nyelvfüggetlen. Emiatt elvileg a módszer valóban hasznos és hatékony, főleg, ha sok nyelvpárra akarjuk elkészíteni a fordítót. Jobb minőségű (valódi fordításra használható, adattal feltöltött) rendszert még nem láttam.

9.4 Hibrid megoldások a fordításoknál

Mivel messze nincs jó minőségű fordító, elvárható lenne, hogy több fordító összedolgozzon. A gond ott van, hogy túl nagy választék nincs a gépi fordítókból. Ha statisztikai módszerekkel dolgozó gépi fordításokat vetünk össze, akkor nagy különbséget nem találunk. Ha más alapon működőket vizsgálunk, akkor meg nagyon nagy különbségek vannak, egy-egy mondat fordítása lényegesen eltér. De ez így van emberi fordításnál is. Minek alapján lehet értékelni, melyik fordítás a jobb? Két fordítás összefésülése vagy lehetetlen, mert annyira eltérnek, vagy felesleges, mert annyira megegyeznek.

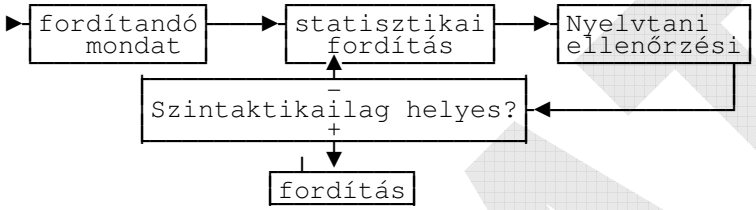
Tulajdonképpen az lenne a cél, hogy két lényegesen eltérő módszert kéne együttesen használni annak érdekében, hogy kiküszöböljék egymás hiányosságait.

A statisztika nagyon jó arra, hogy a tényleges használat felé terelje a megoldást. A generatív, szabályalapú módszerek ezzel szemben a mondatok teljes szerkezetét jobban figyelembe veszik. Ha a két módszert háza-

sítani lehet, akkor annak csak hasznát láthatjuk. Annyira eltér egymástól a két módszerhez használt eszköz, hogy nem nagyon lehet egymással összefonódva egy algoritmussá kovácsolni.

Szabály alapú rendszereket ugyan kiegészíthetnek statisztikából eredő információval, de ezeket közvetlen a szabályokhoz szokták rögzíteni. Emiatt inkább az a megoldást érdemes választani, hogy az egyik módszer alkalmazása után a másik elv alapján szelektálnak. Ha például szabály alapú fordítás több eredményt tart esélyesnek, akkor a két eredményt a statisztika alapján kiértékelik.

Ennek a megfordítottja a gyakoribb. Az előzőekben tárgyalt shake and bake módszert itt is lehet alkalmazni. A lényege az, hogy statisztikai alapon nyert eredményt generatív elven működő elemzőnek adnak át, és ha ez elfogadja, akkor a megoldást esélyesnek tartjuk. Ha a szintaxis próbáját nem állja ki, akkor elvetjük, még akkor is, ha a valószínűség rá szavaz.



A statisztika rész vonatkozhat egyes részek kiértékelésére, de lehet szórendi „keverés” is.

9.5 Szövegszinkronizáció

Szövegszinkronizáció

10. Alkalmazások

10.1

10.2 Stiláris és mondatszintű ellenőrzés

10.3 Szövegelemzés

10.4 Szövegkeresés, indexelés

10.5 Intelligens helyettesítés

10.6 Szótárak

10.7 Gépi és géppel segített fordítás

10.7.1 Szintaktikus elemzésen és generáláson alapuló fordítás

10.7.2 Fordító memóriák

10.7.3 Szótár alapú fordítás

10.7.4 Hibrid rendszerek

10.7.5 Egy magvú és nyelvpárokra épülő rendszerek

10.8 Fordítástámogatás konzisztencia vizsgálat – kontrollált nyelvek.

Irodalomjegyzék

- Kiefer Ferenc—Ábrahám Samu: A full-fledged model of machine translation. Kybernetika 1 (1965) 348—364.
- Kiefer Ferenc—Ábrahám Samu: Some remarks on linguistic theory. Acta Linguistica Hung. 15 (1965) 287—295.
- Kiefer Ferenc— Ábrahám Samu: About the formalization in linguistics. Linguistics 17 (1965) 11—20.

1. David R. Dowty, Lauri Karttunen & Arnold M. Zwicky: Natural Language Parsing. Cambridge University Press, 1985.
2. Prószéky Gábor: Számítógépes nyelvészet. Bp., 1989.
3. Papp Ferenc: A magyar főnév paradigmatikus rendszere Bp., 1975.
4. Thomas Rindflesh: Local ambiguity and natural language processing. In: Mushira Eid & Gregory Iverson (eds.): Principle Analysis of Natural Language (Current Issues in Linguistic Theory, 98). Amsterdam, John Benjamin Company, 1993.
5. Általános Nyelvészeti Tanulmányok I--XVII. Bp.
6. Antal László (szerk.): Modern nyelvelméleti szöveggyűjtemény I--VI. Bp., 1982--1986.
7. Chomsky, N.: Lectures on Government and Binding. Dordrecht, 1981. Foris.
8. É. Kiss Katalin: A magyar mondatszerkezet generatív leírása. Nytud. Ért. 116. sz. Bp., 1983.
9. Levin, Beth-Rappaport, Malka-Zaenen, Annie eds: Papers in Lexical Functional Grammar. Bloomington, 1983. Ind. Univ.
10. É. Kiss, K.: Configurationality in Hungarian. Bp., 1987. Akad. K.
11. Chomsky, N.: Barriers. Cambridge, Mass, 1986. MIT Press.
12. Radford, A.: Transformational Syntax. Cambridge, 1988. Univ.Press.
13. Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Bp., 1992. Akad. K.
14. Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 2. Fonológia. Bp., 1994. Akad. K.
15. Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 3. Morfológia. Bp., 1994. Akad. K.