TERMÉSZETES NYELVEK FELDOLGOZÁSA

Szótárak – lexikonok – hatékony számítógépes tárolása

- **Tárolandó elemek:**
  - 1..43 karakteres szavak, túlnyomórészt nem hosszabbak, mint 17 karakter, átlagosan kb. 10 karakteresek.
  - A karakterkészlet egy természetes nyelv írásjeleiből áll, kb. 116 féle.
- **Követelmények** statikus és dinamikus szótáraknál egyaránt:
  - kis tárigény
  - gyors keresés.
- **További követelmények** dinamikus szótáraknál:
  - gyors beillesztés
  - gyors törlés.
- **Lehetséges szótár-adatszerkezetek:**
  - Nem hatékonyak: lista szekvenciális kereséssel,
    rendezett lista bináris kereséssel
  - Hatékony megoldás: keresőfa, ill. véges állapotú automata.

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

dr.Dudás László

- **Szótárak tárolása keresőfával, vagy véges állapotú automatával**

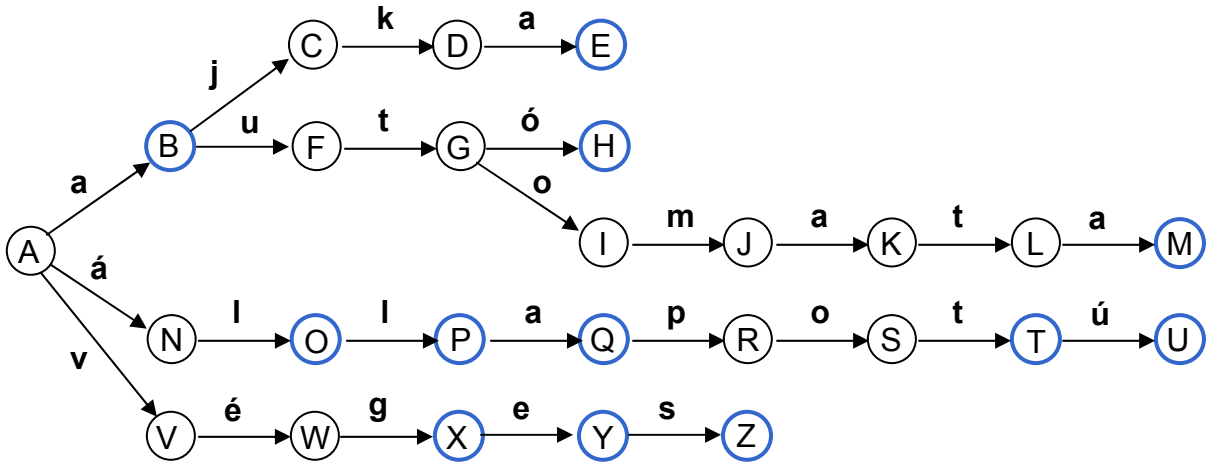
  - A véges állapotú automata (finite state automaton, FSA) **hálós gráf**fal ábrázolható, mely kibontható a kevésbé hatékony memóriafoglalású keresőfának megfelelő **fagráf** alakba.

  - **Előny:** a szó hosszának megfelelő lépésben képesek megállapítani, hogy a szó a szótár része, vagy sem.

  - **Működés:** a bemenetként kapott szó karaktereit sorra véve, a már beolvasott karakterek által meghatározott állapotból egy lépésben elérhető állapotok közül az éppen kapott karakter által meghatározott újabb állapotba lépnek át. Amennyiben az utolsó karakter ún. elfogadó állapotba juttatja, akkor a szót elfogadva állnak le, egyébként a szót nem elfogadva állnak le, azaz a szó nem tartozik a fa, ill az automata által ismert szavak közé.

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999
Képforrás: www.contemplativemind.org

- **Szótárak tárolása keresőfával, vagy végesállapotú automatával ..**
  - **Példaként** adjunk meg egy véges állapotú automatát állapotátmenet-gráffal:



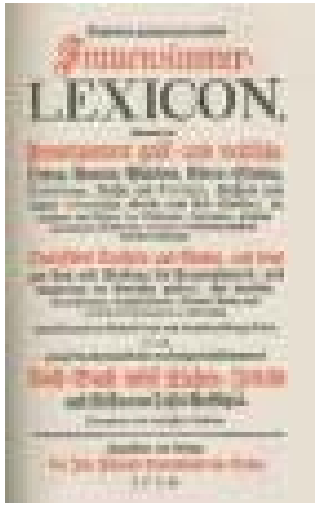
  - Az állapotokat a körökbe írt nagybetűk azonosítják.
  - A szótárban található szavakat elfogadó állapotok vastag körrel jelölve.
  - Az elfogadott szavak: **a, ajka, autó, automata, ál, áll, állapot, állapotú, vég, vége, véges**.
  - Nem elfogadott szavak pl.: **baj, aj, ajkain**.

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

- **Szótárak tárolása keresőfával, vagy végesállapotú automatával ..**

  - Látható, hogy az automata **tárigénye csökkenthető**, ha lemondunk a fagráf-szerkezetről, pl. az E állapotot törölve és D-t L-lel összevonva.

  - Automatás tárolás esetén valós természetes nyelvek esetén a tárigény a listás tárigény **tizedére** csökkenhet.



Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999
Képforrás: www.elmar-lorey.de

Természetes nyelvek feldolgozása: szóelemzés

Természetes nyelvek:

- **Mondatépítés eszközei:** kötött szórend
  elöljárószavak
  toldalékolás.

- **A mondatépítési mód kihatása** a nyelv szavainak sokszínűségére:
  - Nem toldalékoló nyelvek (pl. angol): ~500 000 szóalak
  - Toldalékoló (agglutináló) nyelvek (pl. magyar): ~500 000 000 szóalak.

  meg-szent-ség-telen-ít-het-etlen-ség-es-kedés-e-i-tek-et

- **Morféma:** - tőmorféma = alapszó (szótő, az a rész, mely leginkább meghatározza a toldalékolt szó jelentését)
  - toldalékmorféma, affixum (prefixum: előtag, igekötő; szuffixum: képző, jel, rag).

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

dr.Dudás László

Természetes nyelvek feldolgozása: szóelemzés ..

- **Szöveg morfológiai elemzése:** a szöveg morfémákra **bontása** és a morfématípusok **beazonosítása**. Szükség esetén a morféma alap (lexikális) alakjának meghatározása.

  Pl.:
  labdákat → labdá = labda [főnév] **+** k [többesszám jele] **+** at [tárgyrag]

- **Morfoszintaktikai szabályok:** megadják a morfémák kapcsolódásának szabályait. Gond: a nyelvészeti és számítógépes nyelvészeti optimális megoldások eltérése.

- **A morfológiai elemzés céljai:**
  - a kezelendő szóalakok számának **redukálása** (a toldalékmorfémák száma százas nagyságrendű csak)
  - a mondat szintaktikai, **nyelvtani elemzése** (főnév?, melléknév?, ige?, birtokosjel?, tárgyrag?, stb.)

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

Természetes nyelvek feldolgozása: szóelemzés ..

- **A természetes nyelvek szókészletének implicit tárolása**:
  - szóösszetételek generálásának szabályaival +
  - kivételek kezelésével.

- **A módszer használható:**
  - szintaktikailag (nyelvtanilag) helyes (összetett) **szavak generálására**
  - megadott szavak szintaktikai **helyességének ellenőrzésére**.

- **A módszer hibái:**
  - **Túlgenerálás** – a kivételkezelés hiányosságai és a jelentés tárolásának hiánya miatt olyan szavakat is létrehoznak, illetve elfogadnak, amelyek nincsenek az élő nyelvben: pl. almavaj
  - Inkább zártak, mint nyíltak: a nyelvben megjelenő **új szavak kezelése** nem automatikus.
    Fokozatok a nyitottságban:
    - nyitott új **szó**ra (gyakori igény)
    - nyitott új nyelvtani **szabály**ra (közepesen gyakori)
    - nyitott új **toldalék** megjelenésére (igen ritka).

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

Morfológia típusok

A szóösszetételek, a helyes morfémaláncolatok képzésének megadására alkalmas szabályokat megadó modellek típusai.

1. **Kétszintes morfológiák** (Koskenniemi, 1983):
   A felbontott szó morfémájához megadja az (esetleg kissé eltérő) alap (lexikális) **morfémát** a morféma **nyelvtani kategóriájával együtt**. Pl.:
   labdá+k → labda [főnév] + k [többesszám jele].
   A karakterről karakterre haladó elemzés véges automatát alkalmaz és jellemzője, hogy **megfordítható**: pl.:
   labda [főnév] + k [többesszám jele] → labdá+k .
2. **Folytatási osztályok:** egy morfémához megadja a lehetséges folytatómorfémákat. Pl.:
   labda [főnév] **(**+t [tárgyrag], +val [eszközhatározó rag], +nak [birtokosrag], … **);**
   +k [többesszám jele] (+at[tárgyrag], +nak [birtokosrag], +val [eszközhatározó rag],+ból [helyhatározó rag] …)
   A morfémák osztályozhatók az egyes folytatási osztályok tartalma alapján.
3. **Unifikációs modellek:** nem a kapcsolódó morfémákat, hanem a morfémát **megelőző és követő morfémák jellemzőit** (morfoszintaktikai és fonológiai, hangképzési tulajdonságok) tárolja. Igekötőknél a megelőző, ragoknál a következő morféma hiányzik, így jellemzői is hiányoznak.

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

dr.Dudás László

Kétszintes morfológia

- Az egyszerű véges automata csak a szintaktikai elemzés legelső lépését képes elvégezni egy toldalékolt szóval: megállapítja, hogy része a modellezett nyelvnek, vagy sem. (Jól írt szó – hibásan írt szó.)

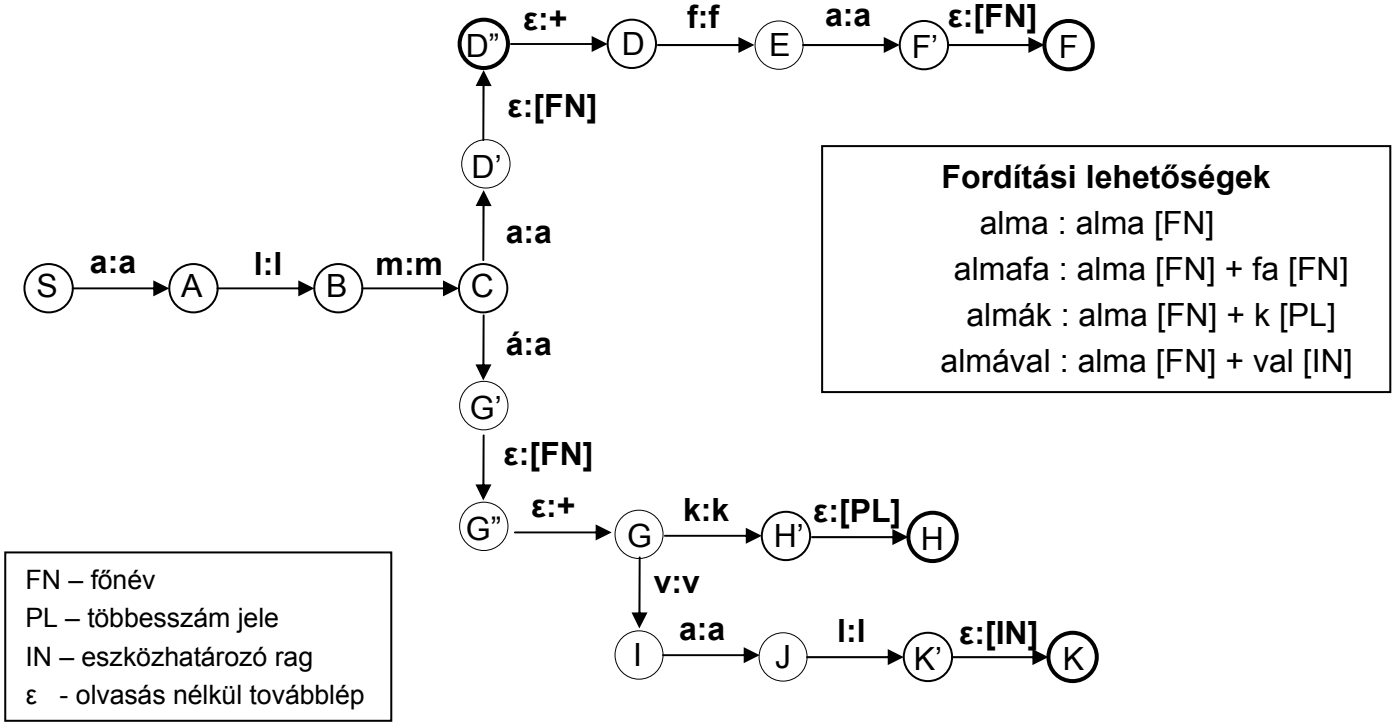
- **Az automata kibővítésével** az automata kimenete gazdagítható: képes megadni
  - a morfémák alap lexikális alakját (értelm -> értelem),
  - a morfémák morfoszintaktikai kódját (-ja -> [birtokos jelző])
  - stb.

- Ilyenkor az automata az input szöveghez **egy output leírást ad**, azaz az inputot lefordítja az outputra. A szöveget (surface level) és a hozzátartozó leírást (lexical level) az automata által kezelt két szintnek tekinthetjük, mivel még az is jellemzi, hogy a leírás is szolgálhat inputként és ekkor az automata azt a szövegre fordítja. Az ilyen, két szint közötti transzformációt végző automatát **véges fordító**nak (finite state transducer, FST) nevezzük.

Felhasznált irodalom: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

Kétszintes morfológia ..

- **Példa***:



- Bár a **visszafelé fordítás** első ránézésre nem egyértelmű, a C állapot után mindkét irányba folytatva a fordítást, a D, ill a G állapotok után a választás egyértelmű.

** Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999*

- **Unifikációs morfológia** (Prószéky, 1994)

  - **Szóelemzés, morfémákra bontás:**
    - morfémák szótári keresésével, plusz
    - a szomszédos morfémák illeszkedésének vizsgálatával.

    Ezáltal a keresés és az illeszkedési szabályok alkalmazása különválik.

  - Egy morfémához le van tárolva a megelőző morfémákhoz való kapcsolódás szükséges feltételeit megadó és a rákövetkező morfémához való kapcsolódás feltételeit megadó ismeret. Kivétel: szókezdő és szózáró morfémák, ahol csak a rákövetkezőhöz, ill. a megelőzőhöz való kapcsolódás létezhet.

  - Példa*:
    szó [ ]
    [+névszó +fn +szótári –elöl –kerek –PL –PERS +ACC –ACCkötő +DAT +INS:V]

** Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999*

- **Unifikációs morfológia** ..

szav [ ] amire a megelőzőnek mind fogadókésznek kell lennie
[+névszó +fn -szótári –elöl –kerek +PL +PLkötő +PERS -ACC +DAT -INS] amit el tud fogadni a rákövetkezőnél, nem kell mind

képez [ ]
[-névszó +szótári +elöl –kerek –ÁS]

képz [ ]
[-névszó +szótári +elöl –kerek –ÁS]

nak [+névszó –elöl +DAT]
[ ]

val [+névszó –elöl +INS:V]
[ ]

kal [+névszó –elöl +INS:K]
[ ]

ak [+névszó –elöl –kerek +PL +PLkötő]
[+névszó –elöl –kerek –PL –PERS +ACC +ACCkötő +DAT +INS:K]

at [+névszó –elöl –kerek +ACC +ACCkötő]
[ ]

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

Unifikációs morfológia ..

A + a morfoszintaktikai-fonológiai tulajdonság meglétét, a – a hiányát jelzi. Pl.:

+fn: főnév	
+szótári : torzítatlan, szótári alapalak	-szótári : nem ilyen
+elöl: hangképzése a szájban elöl történik	-elöl: nem elöl képzett
+kerek: ajakkerekítéses	-kerek: nem ajakkerekítéses
+PL: többesszám lehet	-PL: többesszám nem lehet
+ACC: tárgyrag (-t) lehet	
+ACCkötő: tárgyrag kötőhanggal (-at, -et) lehet	-ACCkötő: kötőhanggal nem lehet
+DAT: részeshatározó rag (-nak, -nek)	
+INS: V v betoldás lehet	-INS: nem lehet betoldás
+ÁS: -ás, -és képző lehet	-ÁS: nem lehet
(- PERS: a forrásban nem definiált)	

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

Unifikációs morfológia ..

- A morfológiai elemzés, azaz a megfelelő morfémák egymásra-következésének meghatározásához két dolgot kell nézni:

  1. a morfémák szón belüli sorrendjét

  2. a szomszédos morfémák egymáshoz illeszkedését a kapcsolódási pontban: a rákövetkező morféma igényeit a megelőző morfémának teljesítenie kell.

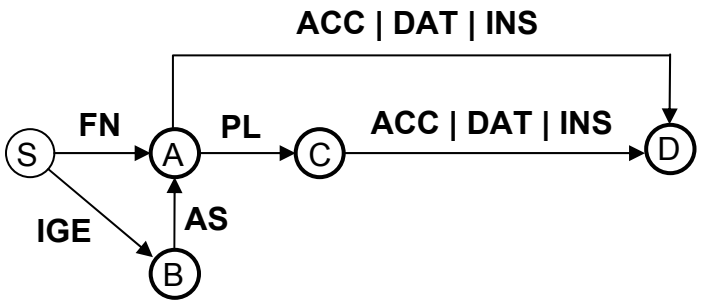
Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

- **Unifikációs morfológia** ..

1. **A morfémák szón belüli sorrendjének meghatározása**

   A morfémák lehetséges sorrendjei modellezhetők véges automatával, melynek átmeneti **operátorai most az egyes morfématípusok**. Egy lehetséges morfématípus-kapcsolódási sorozat az automata egy bejárási útját adja.

   Pl.:



Az automata például a következő morfématípus-sorozatokat fogadja el:
FN+PL (ház+ak);
FN+PL+ACC (ház+ak+at);
IGE+AS+DAT (vág+ás+nak)
de a következőket nem:
FN+PL+AS (ház+ak+ás);
FN+DAT+PL (ház+nak+ok)

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

- **Unifikációs morfológia** ..

2. Illeszkedésvizsgálat, unifikációval

Az 1. lépés még nem ad minden esetben jó megoldást. Ezért a második lépéssel **ki kell zárni a helytelen elfogadásokat** a kapcsolódó morfémák morfoszintaktikai-fonológiai elvárásainak egyeztetésével. A kapcsolódás feltétele, hogy a két jegyszerkezetben, elvárássorban **ne legyenek azonos jellemzők eltérő értékkel** (előjellel)
Pl. a következő szavak felbontása, morfológiai felbontása **helyes**nek vehető a vastagon jelzett elvárások egyezése miatt:

szó
[+névszó +fn **+szótári** –elöl –kerek –PL –PERS +ACC
–ACCkötő +DAT +INS:V]

szó+nak
[**+névszó** +fn +szótári **–elöl** –kerek –PL –PERS +ACC
–ACCkötő **+DAT** +INS:V]
[**+névszó –elöl +DAT**]

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

- **Unifikációs morfológia** ..

szó+val
[**+névszó** +fn +szótári **–elöl** –kerek –PL –PERS +ACC
–ACCkötő +DAT **+INS:V**]
[**+névszó –elöl +INS:V**]

szav+ak+at
[**+névszó** +fn -szótári **–elöl –kerek +PL +PLkötő** +PERS -ACC +DAT -INS]
[**+névszó –elöl –kerek +PL +PLkötő**]
[**+névszó –elöl –kerek** –PL –PERS **+ACC +ACCkötő** +DAT +INS:K]
[**+névszó –elöl –kerek +ACC +ACCkötő**]

szav+ak+kal
[**+névszó** +fn -szótári **–elöl –kerek +PL +PLkötő** +PERS -ACC +DAT -INS]
[**+névszó –elöl –kerek +PL +PLkötő**]
[**+névszó –elöl** –kerek –PL –PERS +ACC +ACCkötő +DAT **+INS:K**]
[**+névszó –elöl +INS:K**]

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999

- **Unifikációs morfológia ..**

  - **Példák az automata által elfogadott, de az unifikáció által a pirossal jelzett ütközések alapján kiszűrt esetekre:**

    szav
    [+névszó +fn **-szótári** –elöl –kerek +PL +PLkötő +PERS -ACC +DAT -INS]

    szav+val
    [**+névszó** +fn -szótári **–elöl** –kerek +PL +PLkötő +PERS -ACC +DAT **-INS**]
    [**+névszó –elöl +INS:V**]

Forrás: Prószéky Gábor - Kis Balázs: Számítógéppel emberi nyelven SZAK Kiadó 1999