A diák összeállításában közreműködött:
Babarczy Anna
Ladányi Enikő

Nyelvtechnológia

Balázs Andrea

Látás, nyelv, emlékezet

Mire jó a nyelvtechnológia?

- Helyesírás-ellenőrző
- Beszédfelismerés
- Gépi fordítás
- Gépi összegzés, szövegkivonatolás
- Jegyrendelés

Propozicionális reprezentáció, mint a nyelvtechnológia alapja

- A világot leképző modellek szimbólumokból építkeznek, amelyek kapcsolata a világ egy reprezentációját adja
- Propozició = állítás (<=> imperatívus)

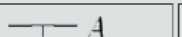
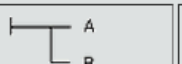
Propozicionális reprezentáció: Minden amit tudunk kijelentések formájában van a fejünkben

Frege a modern propozicionális reprezentáció első kifejtője

- **Frege 1879** *"Fogalomírás, a tiszta gondolkodás formulanyelve az aritmetika mintája szerint"*

- Argumentum-függvény (alany-állítmány) helyett

Frege fogalomírásának alapjelei

Alapfogalom	Frege jelölése	Modern jelölés(ek)
Ítélés	$\vdash A, \models A$	p(**A**) = 1; p(**A**) = i
Tagadás		$\neg A, \sim A$
Feltételesség (kondicionális)		$B \to A$ $B \supset A$
Univerzális kvantifikáció		∀**y**: **Φ**(**y**)
Egzisztenciális kvantifikáció		∃**y**: **Φ**(**y**)
Egyenlőség	$A \equiv B$	$A = B$
A *Fogalomírás*ban még nem bevezetett alapjelek:		
Függvény (értékmenete)	$\alpha \cdot \Phi(\alpha)$	$\mathfrak{a} \mapsto \Phi(\mathfrak{a})$
Deskriptor	$\mathrm{I}\xi \cdot \Phi(\xi)$	$(\mathrm{I}x)\Phi(x)$

```
<!DOCTYPE html>
<html>
    <head>
        <title>Example<title>
        <link rel="stylesheet" href="styl
    </head>
    <body>
        <h1>
            <a href="/">Header</a>
        </h1>
        <nav>
            <a href="one/">One</a>
            <a href="two/">Two</a>
            <a href="three/">Three</a>
        </nav>
```

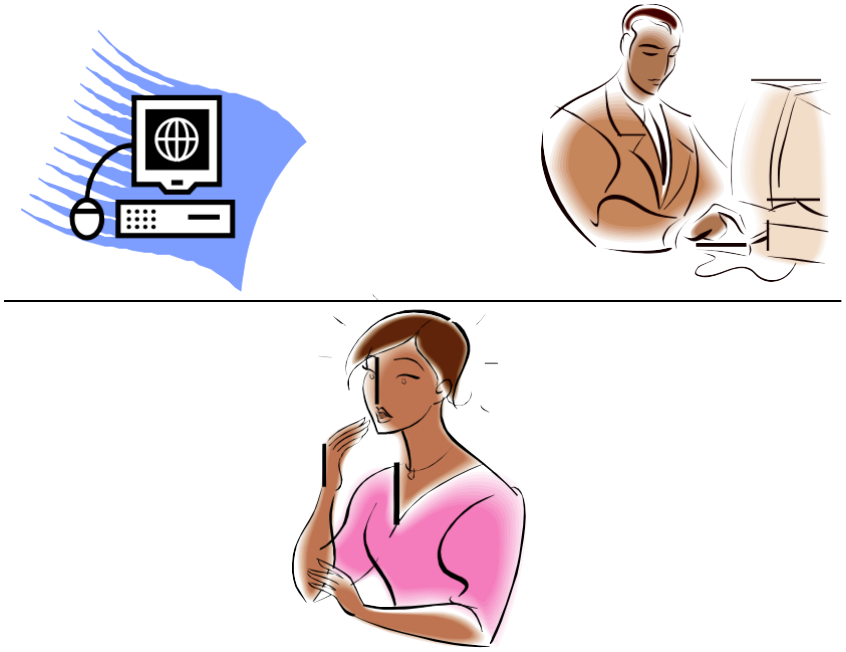
Klasszikus kognitív felfogás

- **Szimbólikus** soklépcsős átkodolási folyamatok
- Az információfeldolgozás **szekvenciális**
- A megimsermés **egyelen aktív feldolgozó központ**ja
- Minden feladatot egyetlen közös **propozicionális** nyelvre fordítunk le
- A feldolgozás **gépies**: kicsiny kapacitású operatív és óriási kapacitású háttértárak

Mesterséges intelligencia
Intelligens-e a gépi társalkodórendszer?

- Alan Turing, 1950: a Turing teszt (gondolatkísérlet)

Akkor megy át egy gép a próbán, ha a középen lévő ember nem tudja eldönteni, géptől vagy embertől jött-e a válasz?



Eliza: „pszichoterapeuta”
Weisenbaum 1966

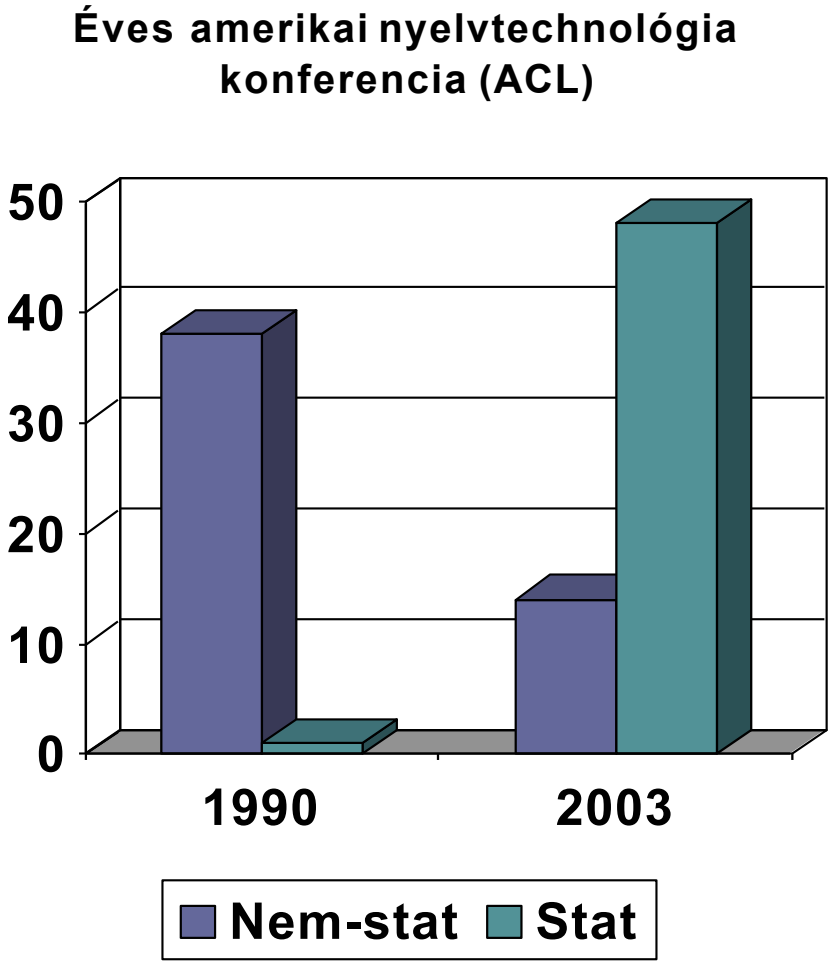
- A férfiak mind egyformák.
- Ezt hogy érted?
- Mindig nyaggatnak minket valamiért.
- Mondanál egy konkrét példát?
- Hát, a fiúm küldött ide.
- A fiúd küldött ide.
- Azt mondja, mindig depressziós vagyok.
- Sajnálom, hogy depressziós vagy.
- Talán ki kellene békülnöm az anyámmal.
- Mesélj még a családodról.

- A módszer:
  - Kulcsszókeresés
    - mindig
    - anyám
  - Kulcsszó válasz-szabályokat aktivál
    - Mondanál egy konkrét példát?
    - Mesélj még a családodról.
  - Reguláris kifejezések segítségével behelyettesítés a szabálynak megfelelően
    - s/.* (depressziós|szomorú) vagyok .* /Sajnálom, hogy \x vagy/

Eliza

Nyelvtechnológia ma: Két alapelv

- Szabályalapú (Nem-stat)
- Példaalapú statisztikai
  - korpuszok
- Két egymással ellentétes cél:
  - Lefedettség növelése (hamis negatívok csökkentése) -- lazítás
  - Pontosság növelése (hamis pozitívok csökkentése) -- szigorítás



Gépi nyelvfeldolgozás általános szintjei

- **Beszédfelismerés (inger érzékelés)**

Bayes, N-gram

- **Parsing (elemzés)**

szófaji, morfológiai, szintaktikai

- **Szemantikai elemzés (értelmezés)**

Gépi nyelvfeldolgozás általános szintjei

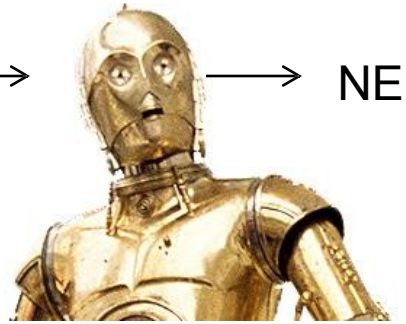
- **Beszédfelismerés (inger észlelés)**

Bayes, N-gram

- **Parsing (elemzés)**

Parsing (szófaji,morfológiai,szintaktikai)

- **Szemantikai elemzés (értelmezés)**

Nemtom →  → NEM TUDOM

Beszédfelismerés

Hangsorokból szavak

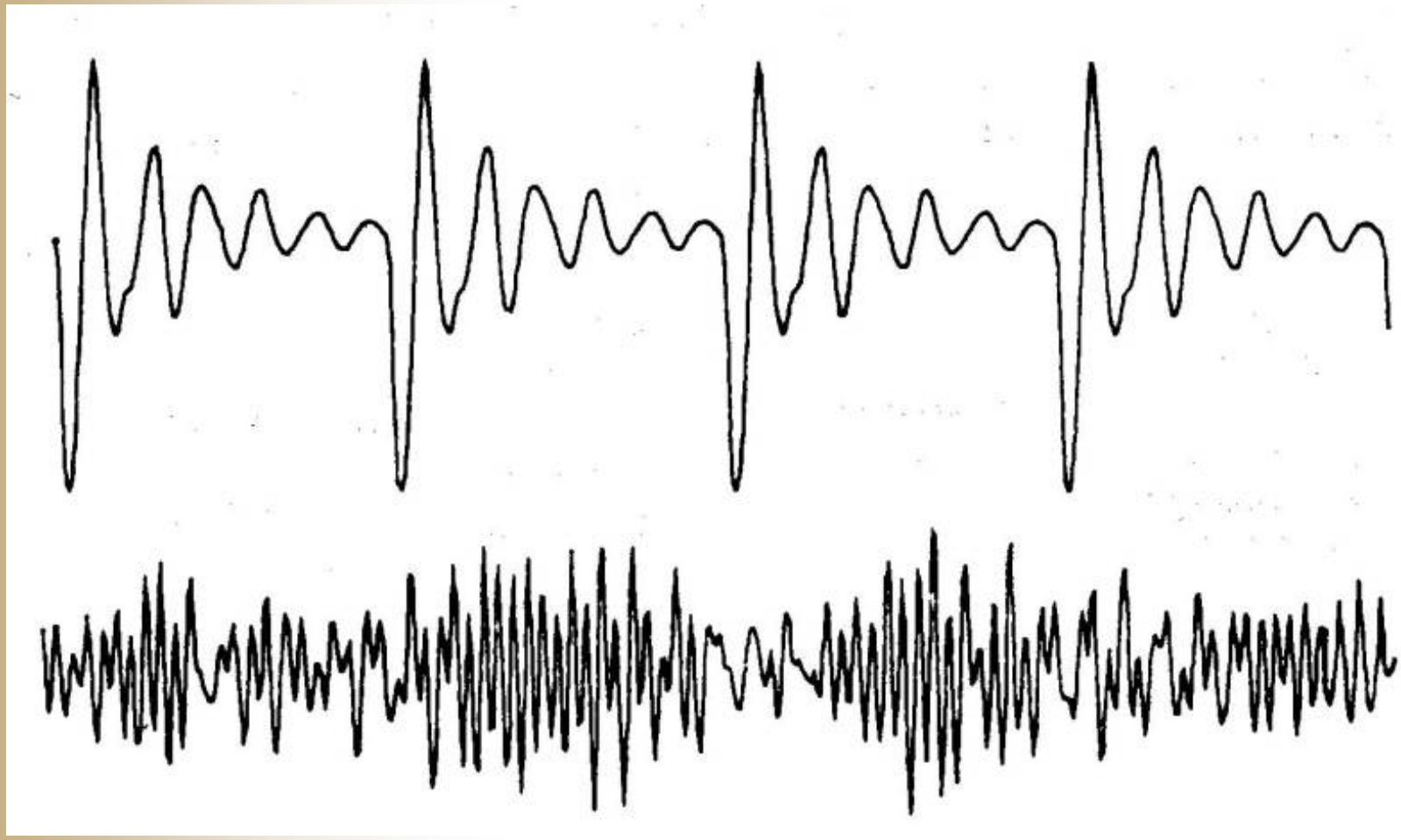
Szükséges tudás

Fonetika: a hangok akusztikai tulajdonságai (formánsfrequenciák)
Fonológia: egy-egy nyelv hangrendszere
Esetleg egy kiejtési szótár

A hang hullámként terjed a levegőben.
A beszédhang összetett, több hullám együtteséből áll (formánsszerkezet).
Kétféle összetett hang:

- periodikus: ismétlődő hullámalak (magánhangzó)
- aperiodikus: rendszeres ismétlődés nélküli (bizonyos mássalhangzók)

[a] és [s] hullámalak

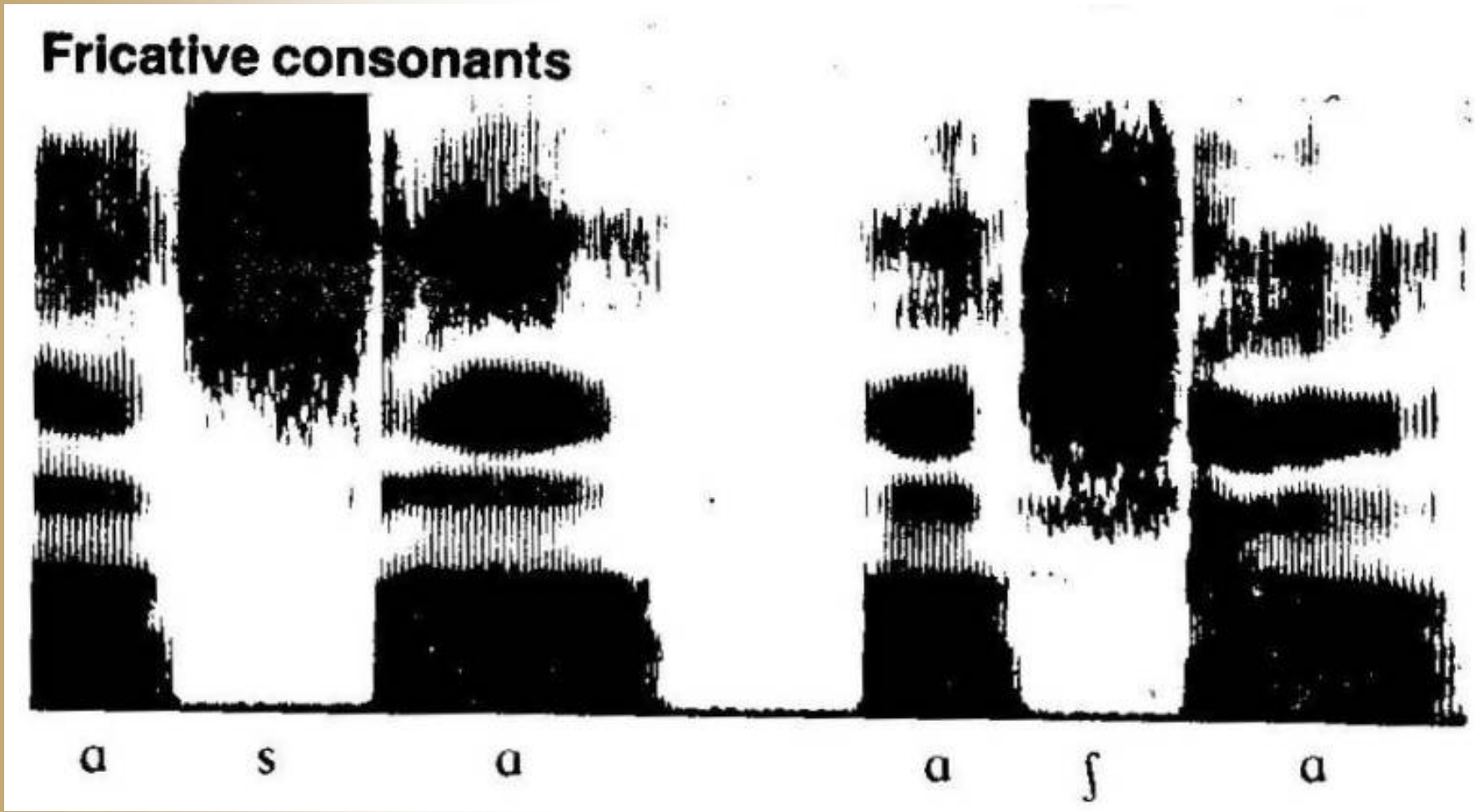


A beszédhang fizikai jellemzői:

- rezgés szaporasága (frekvencia)
- rezgés erőssége (intenzitás)
- rezgés időtartama

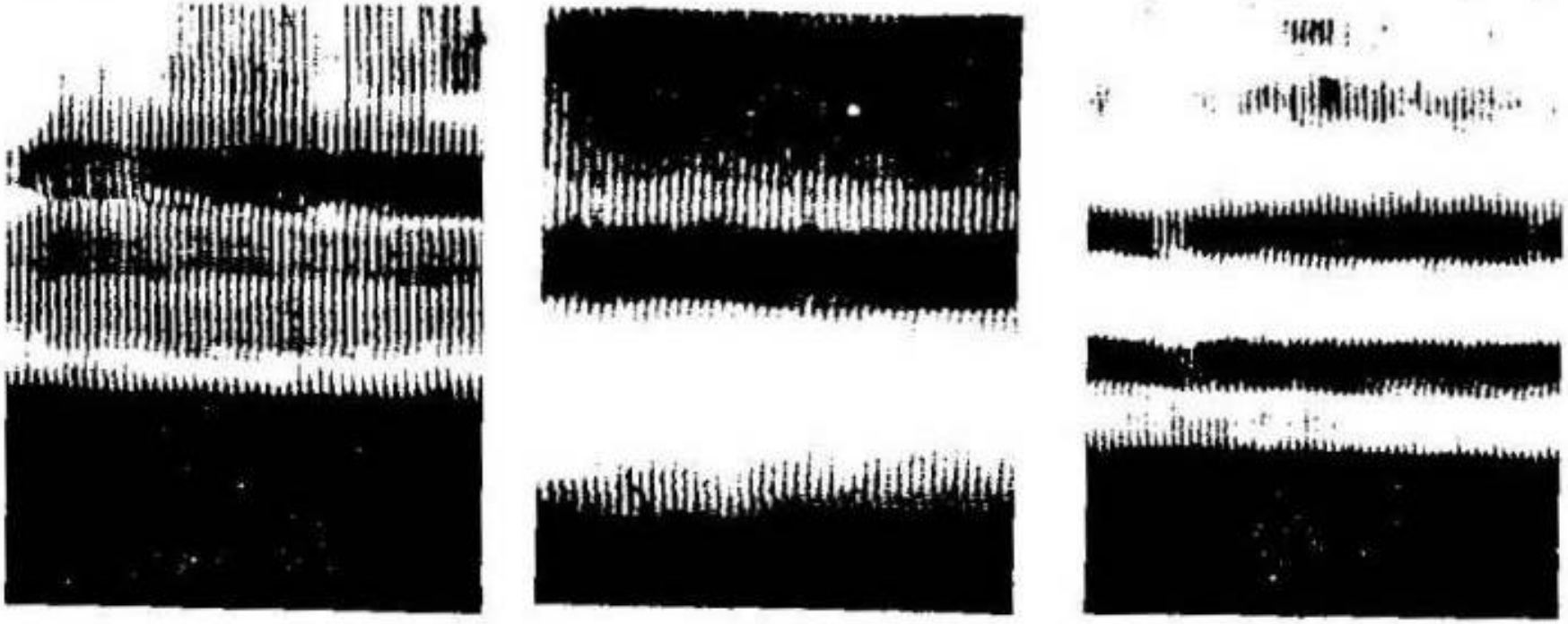
Az összetett hullámok tulajdonságait a spektrogram jeleníti meg:

- vízszintes tengely: idő
- függőleges tengely: rezgésszám (formánsok: F_0 (100-200Hz), F_1 (300-600Hz), F_2 (800-3000Hz))
- harmadik dimenzió (jel erőssége): intenzitás



Gyakori réshangok: f, j, s, sz, v, z, zs

The vowels



Ami bonyolítja a feladatot

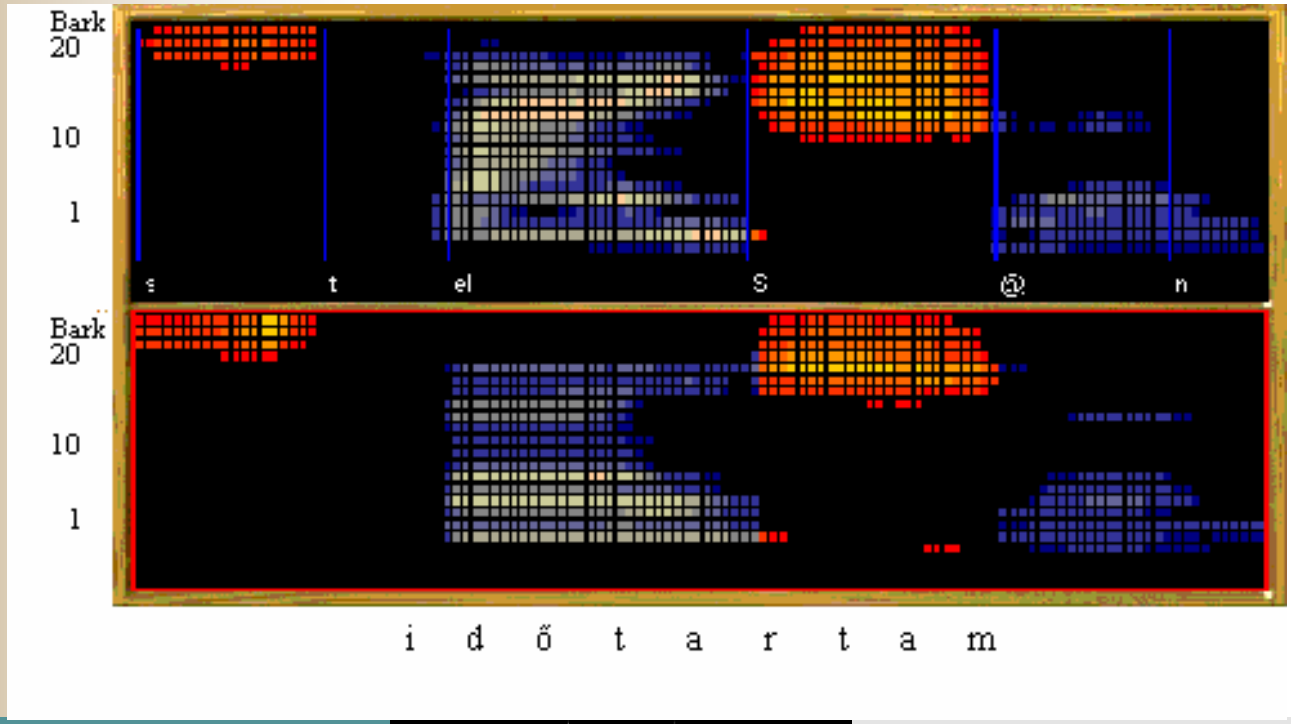
A hangok variabilitása:

Beszélőtől függően (hangerő, hangmagasság, artikulációs különbségek)
Folyamatos beszédben szövegkörnyezettől függően
- Allofónok: n vs. ng
- Hasonulás: mézízű, méztartalmú, mésztartalmú

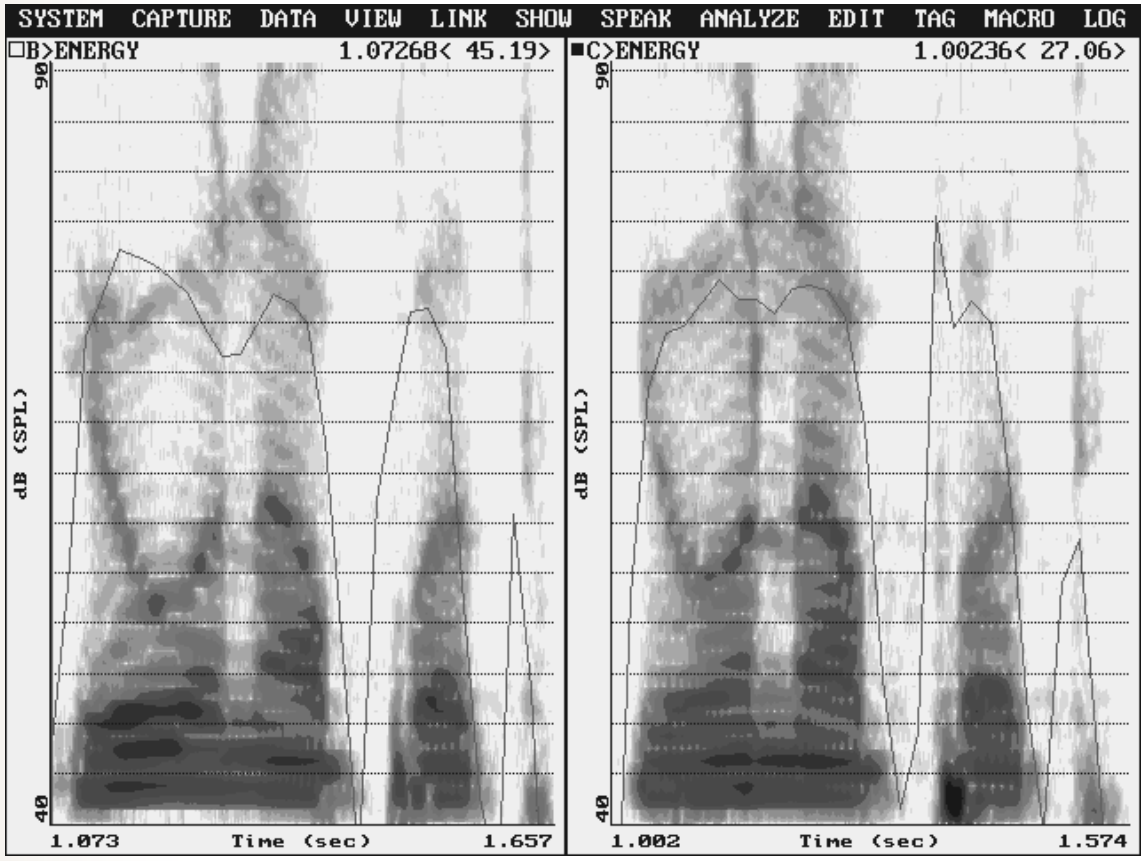
Prozodikai jelenségek: intonáció, hangsúly, ritmus

- A gépi beszédfelismerést nehezíti a folyamatos beszéd zajosság

– *station* ejtése két különböző beszélő által



- *Jó napot!* – ugyanaz a személy egy nap eltéréssel



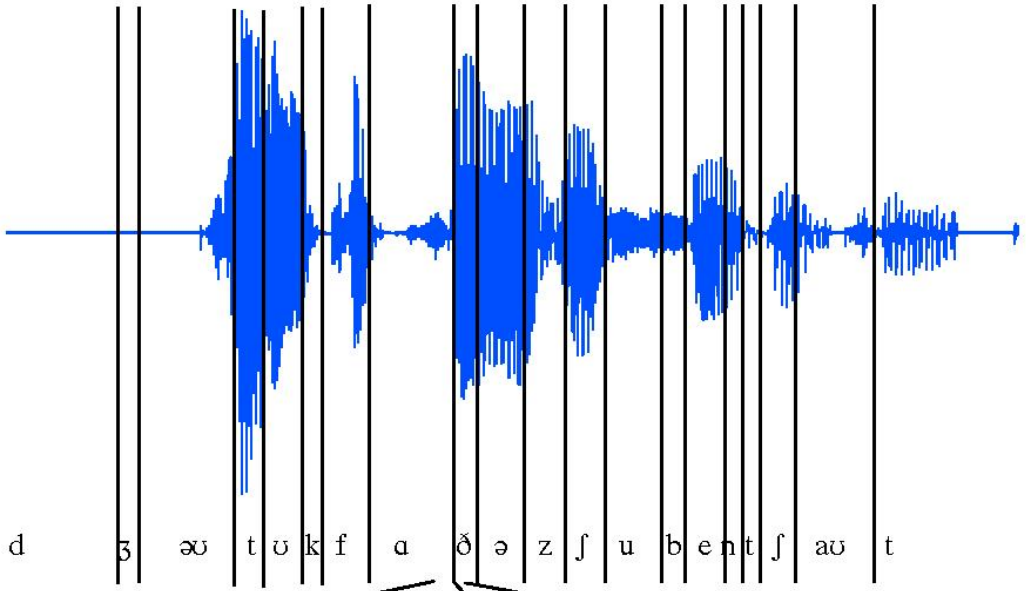
Véges-állapotú transzducer

Az állapotok közötti átmeneteket párok definiálják

Pl. fonéma – **spektrál elemzés** vektor

Minden lehetséges akusztikus jelhez hozzárendel egy vagy több fonémát és megfordítva ({[n], [ng]} ←→ n)

SPECTRAL ELEMZÉS



Label:	ɑ	ɑ	ɑ	ɑ	ð	ð	ð	ð
Frame: 204	205	206	207	208	209	210	211	212
LPC vector:								
2.693137	2.610945	2.597558	2.565878	2.319291	2.349247	2.321899	1.863409	1.481066
–3.15723	–3.07347	–2.94453	–2.91241	–2.11807	–2.11037	–2.02477	–0.63155	0.010306
2.153815	2.379279	1.995438	2.108499	1.147475	0.910962	0.814372	–0.56747	–0.66528
–0.46244	–0.9316	–0.35989	–0.8136	–0.3249	0.256971	0.287573	0.335704	0.133916
–0.62918	–0.35334	–0.83202	0.027484	0.184156	–0.79563	–0.68009	0.429717	0.126124
0.194162	0.306683	0.485961	–0.43509	–0.80504	0.373898	0.156924	–0.89356	–0.28647
0.696667	0.268792	0.402462	1.037625	0.898168	–0.0097	0.262798	0.252545	0.115398
–1.27494	–0.76344	–1.07576	–1.31662	–0.49597	–0.10334	–0.3322	0.661088	0.046145
1.502201	1.145245	1.177903	1.283225	0.157776	0.202759	0.226109	–0.7357	0.109533
–1.37626	–1.13322	–0.6209	–1.15597	0.013327	–0.37699	–0.28064	–0.03584	–0.32924
135756	0.999281	0.119541	1.134817	0.012791	0.677501	0.512114	0.689633	0.540491
–0.52237	–0.65051	0.418474	–0.68714	0.354996	–0.35548	–0.15489	–0.0299	–0.0326
–0.11574	0.121429	–0.70822	0.124007	–0.60873	–0.16315	–0.20147	–0.65855	–0.46588
0.135437	0.046852	0.305646	0.003265	0.24165	0.122075	0.072926	0.305114	0.200167

Egy hasonló transzducer a fonémákat betűkhöz rendelheti

→ beszédfelismerő szoftver

De a módszer nem elég megbízható…
Ki kell egészíteni valószínűségekkel.
„Top-down” módszer segíthet

Bayesi beszédfelismerés (fonéma valószínűégek megállapítása)

Ha adott egy hangsor *h*, mi a valószínűsége egy *s* szónak: P(s|h)

- Korpusz: szógyakoriság
- Variáció-korpusz: hangsorokhoz rendelt szólisták ([tom] -> Tom, tudom, atom)
- Fonotaktikai folyamatok valószínűségei
  (pl. szóközi szótag lehagyása)
- Legvalószínűbb szó: maximális P(h|s)P(s)

Helyesírásellenőrzés hasonló elveken

Jobb eredmény a szövegkörnyezet figyelembevételével

N-gram modellek (Markov láncok)

Véges-állapotú automaton, ahol az átmenetekhez valószínűségeket rendelünk

A mondat következő szavának prediktálása

Korpusz

A nyelvtan: egy szó valószínűsége, ha az előző (egy, kettő, három...) szó a mondatban adott: $P(s_n|s_{n-1})$

s_n gyakorisága s_{n-1} után

Osztva s_{n-1} gyakoriságával

Minél nagyobb a szövegkörnyezet, annál pontosabb a nyelvtan:

Északnyugat felől felszakadozik, csökken a ___

N-gram (biagram)nyelvtan

egy szó valószínűségi előfordulása adott szövegkörnyezetben

	szem	eszem	Azt hiszem
zöldeskék	.059	.000	.000
villával	.000	.013	.000
komolyan	.000	.001	.721

Bigram nyelvtan

	atom	tudom	találom
nem	.000	.470	.022
helyesnek	.000	.001	.009
iráni	.003	.000	.000

Gépi beszédfelismerő rendszer a három tenchnika kokmbinálásával

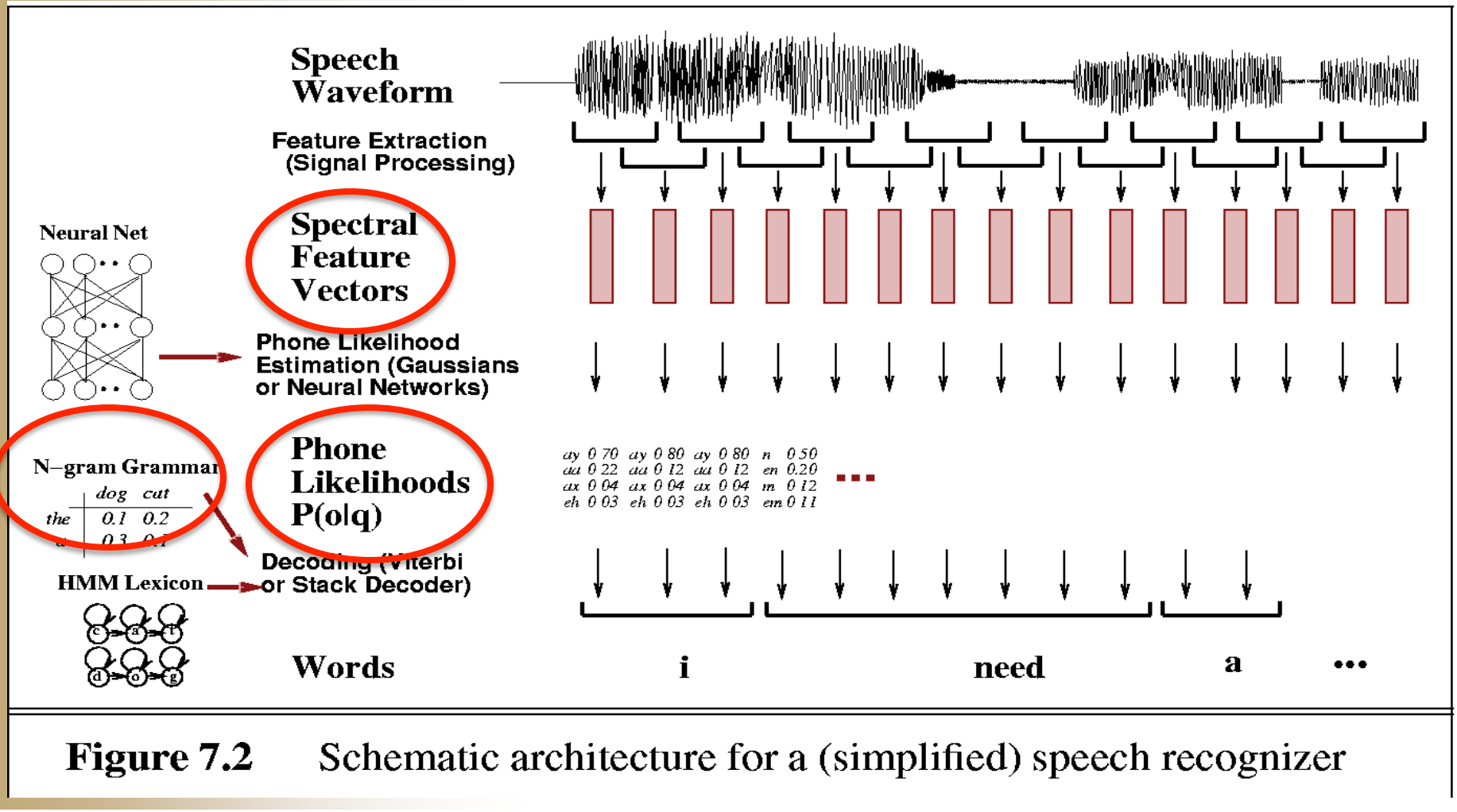


Figure 7.2 Schematic architecture for a (simplified) speech recognizer

Az n-gram modell hátrányai:

- Stílus- és témafüggő
- Hatalmas korpuszokat igényel

Gépi nyelvfeldolgozás általános szintjei

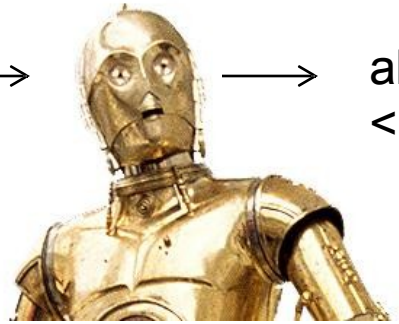
- **Beszédfelismerés (inger érzékelés)**

Bayes, N-gram

- **Parsing (elemzés)**

szófaji, morfológiai, szintaktikai

- **Szemantikai elemzés (értelmezés)**

ALAKÍTOTTÁK ⟶  ⟶ alak<NOUN[ÍT]VERB<PAST><PLUR><DEF>>

Parsing

Szófaji és morfológiai elemzés

Elemző komponensei

- Szófaj meghatározása + morfológiai összetétel elemzése
- Szótár
  - Szótövek listája
    - Szófaji kategóriájuk (főnév, ige, határozószó, stb)
    - Morfofonológiai kategóriájuk
      - Sziszegő végű: olvas (olvasol és nem olvassz)
        <=> (pl.: hallasz, vársz, ugrasz)
    - Alternatív töveik (pl.: bokor, bokr -> bokrot)

-> ELEMZÉS CSAK SZÓTÁRRAL?

Elemző komponensei

- Szabályok
  - Toldalékok listája
  - Milyen szófajhoz milyen morfológiai jegy tartozhat
  - Morfofonológiai kategóriák:
    - Sziszegő végű: olvas, olvasOL, *olvasSZ
    - Egyéb: fél, félSZ
  - Morfotaktika: affixumok sorrendje
    - Kenyer-em-et, *kenyer-et-em

Az elemzés feladatai

- Szóalakok felcímkézése a szótár és nyelvtan szabályai alapján

→ egy egyértelmű eredmény

→ több lehetséges elemzés

→ nem található a szótárban

- alak
  alak<NOUN>
- alakult*1*
  alak<NOUN[UL]VERB<PAST>>
- alakult*2* alak<NOUN[UL]VERB[PAST_PART]ADJ>
- alakították
  alak<NOUN[ÍT] VERB<PAST><PLUR><DEF>>
- alakítsunk
  alak<NOUN[ÍT] VERB<SUBJUNC-IMP><PERS<1>><PLUR>>

Kihívás: A többértelműség feloldása

- Szövegkörnyezet segítségével
- *A tűz felmelegítette az átfagyott túrázókat*
- *tűz <NOUN>*
- *tűz <VERB>*
- *A tűz → tűz <NOUN>*
- *Ma megint erősen tűz a Nap*
- *tűz <NOUN>*
- *tűz <VERB>*
- *tűz a → tűz <VERB>*

Ismeretlen szavak

Ha a szótárban nem szerepel egy szó...

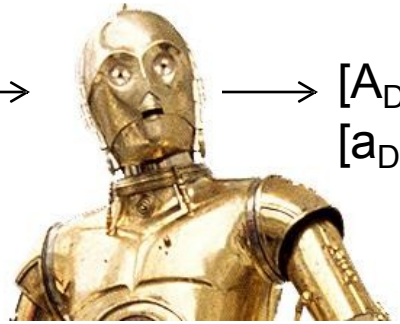
- Többértelmű címkézés+egyértelműsítés szövegkörnyezet alapján
  - egyenletes elosztásban (minden címke)
  - címke-gyakoriság szerint (bizonyos gyakorisági küszöb fölött)
- Morfológiai szerkezet alapján

– Roxfortban Szémisen rottolnak a makánok a leghöntebb mufjotukban.

- 95 – 98% teljesítmény (a címketár méretétől függően)

Megfelel az ember teljesítményének!

A_{Det} $kutya_{NounNom}$ $kergette_{VerbPast}$
a_{Det} $macskát_{NounAcc}$



$[A_{Det}$ $kutya_{Noun}]$ $_{NP}$ $[kergette_{VerbPast}$
$[a_{Det}$ $macskát_{NounAcc}]_{NP}]_{VP}$

Parsing

Mondat szerkezeti elemzése

Mondatelemzés

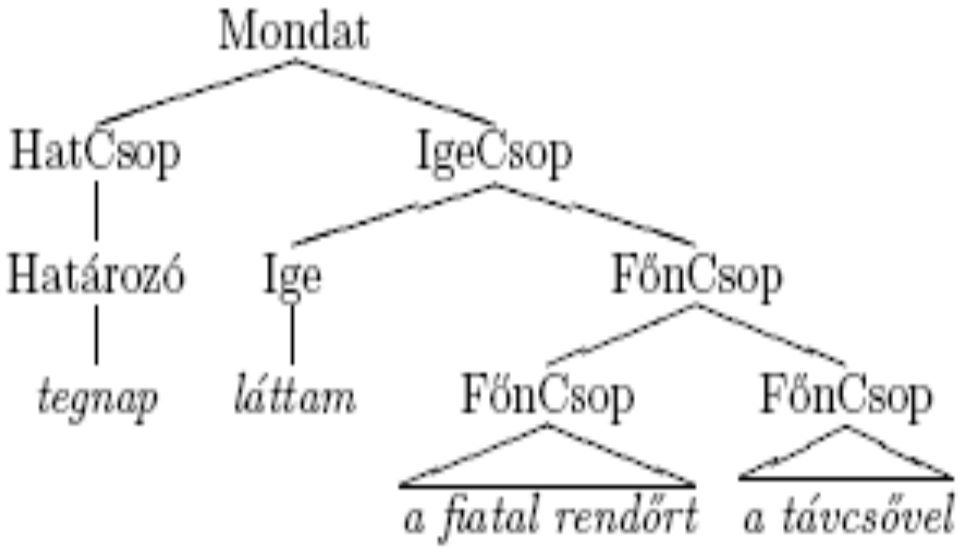
- Elemzés célja: a szavak közötti szintaktikai/szemantikai viszonyok feltárása
  - Dependencia nyelvtanok:

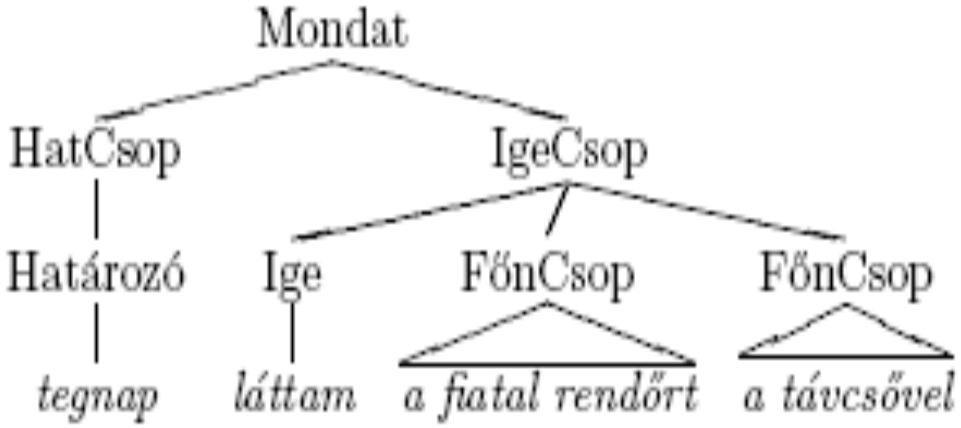
    A nyelvtan kiterjesztett szókapcsolattár
  - Frázis-struktúra nyelvtanok:

    Mondatszerkezet feltárása

    Szerkezeti többértelműségek kimutatása

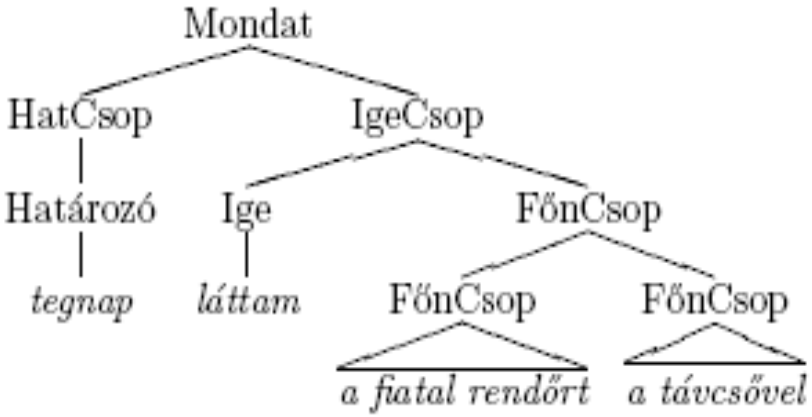
    Hierarchikus rend

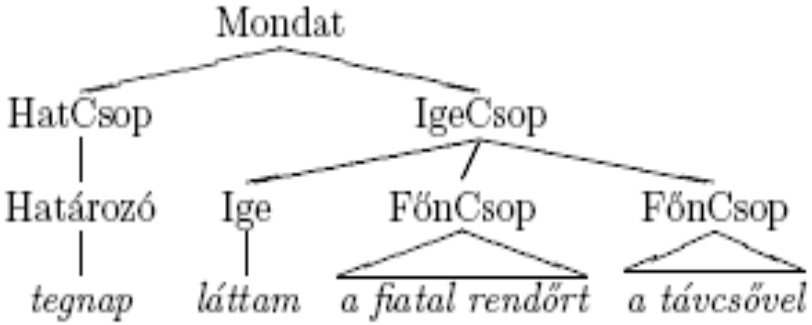

Mondat
HatCsop
IgeCsop
Határozó
Ige
FőnCsop
tegnap
láttam
FőnCsop
FőnCsop
a fiatal rendőrt
a távcsővel


Mondat
HatCsop
IgeCsop
Határozó
Ige
FőnCsop
FőnCsop
tegnap
láttam
a fiatal rendőrt
a távcsővel

Környezet-független nyelvtan

- Mondat → Határozói_fr Igei_fr
- Határozói_fr → Határozószó
- Igei_fr → Ige Főnévi_fr
- Főnévi_fr → Főnévi_fr Főnévi_fr
- Főnévi_fr →
  Névelő (Melléknévi_fr) Főnév
- Főnévi_fr → Névelő Főnév
- Melléknévi_fr → Melléknév
- Határozószó → tegnap
- Ige → láttam
- Névelő → a
- Melléknév → fiatal
- Főnév → rendőrt, távcsővel
- ...





- + Morfológiai megkötések
  - *Tegnap láttalak a rendőrt a távcsővel
  - *Tegnap látom a rendőrt a távcsővel.
  - *Tegnap láttam a rendőrhöz a távcsőnek
- + Szabadabb szórend
  - Láttam tegnap a rendőrt a távcsővel.
  - A rendőrt tegnap láttam a távcsővel.
  - A távcsővel a rendőrt tegnap láttam.

Gépi nyelvfeldolgozás általános szintjei

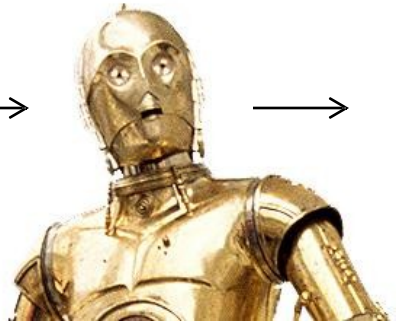
- **Beszédfelismerés (inger érzékelés)**

Bayes, N-gram

- **Parsing**

szófaji, morfológiai, szintaktikai

- **Szemantikai elemzés (értelmezés)**

Van egy autóm ⟶  ⟶ ∃x,y Birtokol(x)∧Birtokló (Beszélő,x)∧ BirtokolValamit (y,x)∧Autó(y)

Szemantika

A jelentés meghatározása

Információk kivonása

- „Mondjon nekem reggeli járatokat kedden Bostonból San Franciscoba.”

MUTAT:

JÁRAT:

EREDET:

VÁROS: Boston

DÁTUM: kedd

IDŐ: reggel

CÉL:

VÁROS: SF

- LISTÁZ -> mondjon nekem | szeretnék | mutatna|…
- INDULÁSIIDŐ -> ÓRA (előtt|körül|után) | reggel | délután | este
- ÓRA -> egy|két|három…|huszonnégy
- JÁRAT -> (egy) járat|járatok
- EREDET -> VÁROS-EREDET_HELYRAG
- CÉL -> VÁROS-CÉL_HELYRAG
- VÁROS -> Boston | San Francisco | Budapest

Mondat tematikai elemzése

- Frázisok → Tematikus szerepek a morfoszintaktikai struktúra alapján

A kutya tegnap a házig kergette a macskát.

- Alany → Ágens
- Tárgy → Páciens
- Helyragos NP/PP/helyhatározó → Cél
- Időhatározó/PP/ragozott NP → Idő

Gépi fordítás

- Szabályalapú rendszerek
  - morfoszintaktikai és szemantikai elemzés
  - nyelv-független általánosítás
  - szöveg generálása a célnyelven
- Statisztikai rendszerek
  - Parallel korpuszok: (fordítóprogramok betanítása)
- A kettő kombinációja

Webforditas.hu

- *Az olvasónak mindenesetre jó találgatást és kevés tévedést kívánnak a szerkesztők.*

  The editors wish the reader a good guessing and few mistakes whatever.

  A szerkesztők kívánnak az olvasó egy jó találgató és kevés hiba bármi.

A google fordító mint példa

- Statisztikai elemzések
- Nagy mennyiségű kétnyelvű szövegkorpuszok
- Öntanuló algoritmus
- Felhasználói visszajelzések figyelembe vétele

Szövegkivonatolás

- Kivonatolás
  - Szavak, szókapcsolatok, mondatok kiválasztása a szövegen belüli gyakoriság és pozíció és az általános gyakoriság alapján
- Absztraktkészítés
  - Jelentésreprezentáció a szövegről, és ez alapján generál összefoglalót

Chatterbot-ok

- ELIZA
- A.L.I.C.E. – többszörös Loebner díjas
- Jabberwacky
- Kyle
- Mitsuku

Köszönöm a figyelmet!