Számítsuk ki a nyelvet!
Matematika, fizika és algoritmusok a nyelvben

Biró Tamás

Eötvös Loránd Tudományegyetem

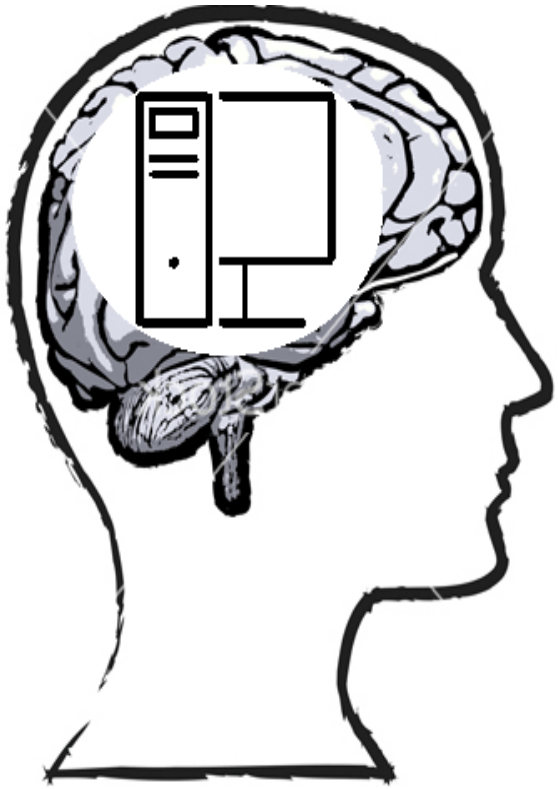
KöMaL Ifjúsági Ankét, 2015. október 28.

Hogyan működik az emberi agyban a nyelvi app?

Mi zajlik az emberi agyban?

Behaviorizmus az 1950-es évekig:
az emberi agy, mint fekete doboz.

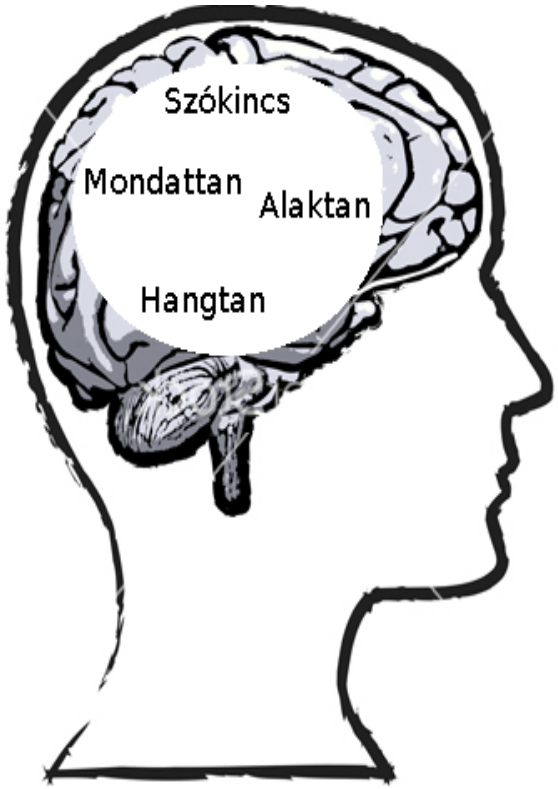
Hogyan működik az emberi agyban a nyelvi app?



Mi zajlik az emberi agyban?

Kognitív fordulat az 1960-as években:
az emberi agy, mint számítógép.

Hogyan működik az emberi agyban a nyelvi app?



Mi zajlik az emberi agyban?

Kognitív fordulat az 1960-as években:
az emberi agy, mint számítógép.

A nyelvészet, mint *reverse engineering*.

Áttekintés

1. Optimalitáselmélet: egy nyelvészeti modell
2. Az Optimalitáselmélet matematizálása
3. Az Optimalitáselmélet implementálása
4. Összefoglalás és tanulságok

Amire ma nem térek ki:
számítógépes nyelvészet, mint nyelvtechnológia.

Áttekintés

Egy (leegyszerűsített) nyelvészeti tipológia

T.f.h. a világ nyelvei három típusba sorolhatók:

- Szóhangsúly az első szótagon (pl. *hókuszpokusz*).
- Szóhangsúly az utolsó szótagon (pl. *hokuszpokúsz*).
- Szóhangsúly az utolsó előtti szótagon (pl. *hokuszpókusz*).

- Nincs nyelv, amelyben a második szótagon (pl. *hokúszpokusz*).

Nulladik közelítésben igaz. Első közelítésben nagyon sok nyelv ennél komplikáltabb hangsúlyrendszerrel rendelkezik. Az igaz, hogy második szótagos nyelv alig létezik.

OT: a tipológia egy lehetséges modellje

1. Bemenet (szó hangsúly nélkül) $\mapsto$ kimenet (szó hangsúllyal).

2. Bevezetünk három függvényt (a.k.a. constraint, megszorítás, korlát):
   - EARLY: hangsúly minél hamarabb.
     = szó eleje és hangsúlyos szótag közti szótagok száma.
   - LATE: hangsúly minél később.
     = hangsúlyos szótag és szó vége közti szótagok száma.
   - NONFINAL: utolsó szótag ne legyen hangsúlyos.
     = az utolsó szótagon lévő hangsúlyok száma (0 vagy 1).

3. Gen generátor-függvény:
   a bemenethez tartozó lehetséges kimenetelek halmaza.

OT: a tipológia egy lehetséges modellje

1. Bemenet (szó hangsúly nélkül) $\mapsto$ kimenet (szó hangsúllyal).

2. Bevezetünk három függvényt (a.k.a. constraint, megszorítás, korlát):
   - EARLY: hangsúly minél hamarabb.
     = szó eleje és hangsúlyos szótag közti szótagok száma.
   - LATE: hangsúly minél később.
     = hangsúlyos szótag és szó vége közti szótagok száma.
   - NONFINAL: utolsó szótag ne legyen hangsúlyos.
     = az utolsó szótagon lévő hangsúlyok száma (0 vagy 1).

3. Gen generátor-függvény:
   a bemenethez tartozó lehetséges kimenetelek halmaza.

OT: a tipológia egy lehetséges modellje

1. Bemenet (szó hangsúly nélkül) $\mapsto$ kimenet (szó hangsúllyal).

2. Bevezetünk három függvényt (a.k.a. constraint, megszorítás, korlát):
   - EARLY: hangsúly minél hamarabb.
     = szó eleje és hangsúlyos szótag közti szótagok száma.
   - LATE: hangsúly minél később.
     = hangsúlyos szótag és szó vége közti szótagok száma.
   - NONFINAL: utolsó szótag ne legyen hangsúlyos.
     = az utolsó szótagon lévő hangsúlyok száma (0 vagy 1).

3. Gen generátor-függvény:
   a bemenethez tartozó lehetséges kimenetelek halmaza.

OT: a tipológia egy lehetséges modellje

$\mathrm{Gen}(\sigma\sigma\sigma\sigma) = \{$[suuu], [usuu], [uusu], [uuus]$\}$.

/σσσσ/	EARLY	LATE	NONFINAL
☞ [s u u u]	0	3	0
[u s u u]	1!	2	0
[u u s u]	2!	1	0
[u u u s]	3!	0	1

$\mathrm{SF}(\sigma\sigma\sigma\sigma) =$[suuu]

OT: a tipológia egy lehetséges modellje

$\mathrm{Gen}(\sigma\sigma\sigma\sigma) = \{$[suuu], [usuu], [uusu], [uuus]$\}$.

/σσσσ/	LATE	EARLY	NONFINAL
[s u u u]	3!	0	0
[u s u u]	2!	1	0
[u u s u]	1!	2	0
☞ [u u u s]	0	3	1

$\mathrm{SF}(\sigma\sigma\sigma\sigma) =$[uuus]

OT: a tipológia egy lehetséges modellje

$\mathrm{Gen}(\sigma\sigma\sigma\sigma) = \{$[suuu], [usuu], [uusu], [uuus]$\}$.

/σσσσ/	NonFinal	Late	Early
[s u u u]	0	3!	0
[u s u u]	0	2!	1
☞ [u u s u]	0	1	2
[u u u s]	1!	0	3

$\mathrm{SF}(\sigma\sigma\sigma\sigma) =$[uusu]

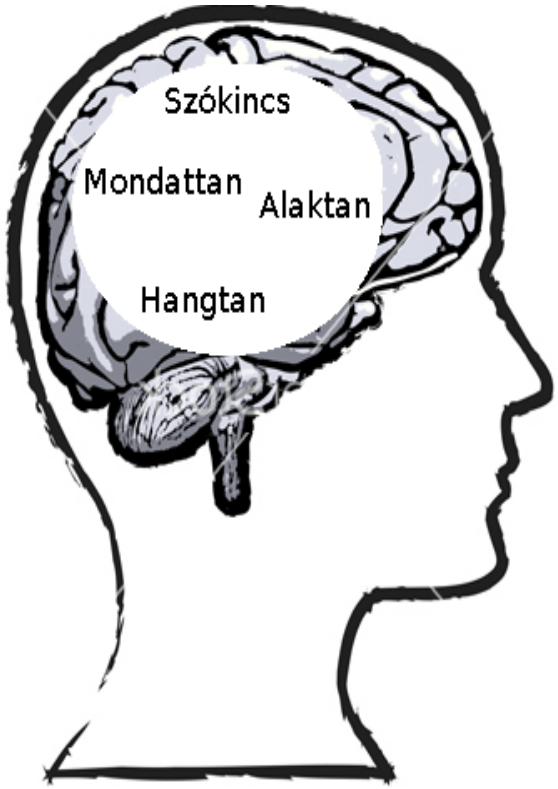
OT: a tipológia egy lehetséges modellje

Az Optimalitáselmélet helyesen **modellezi** a nyelvtipológiát:

- Szóhangsúly az első szótagon:
  EARLY ≫ LATE, NONFINAL, valamint
  NONFINAL ≫ EARLY ≫ LATE
- Szóhangsúly az utolsó szótagon:
  LATE ≫ EARLY, NONFINAL
- Szóhangsúly az utolsó előtti szótagon:
  NONFINAL ≫ LATE ≫ EARLY
- Nincs nyelv, amelyben a második szótagon:
  A három függvény hat permutációja közül
  egyik sem állítja elő ezt a nyelvtípust.

Áttekintés

Hogyan működik az emberi agyban a nyelvi app?



$\mathcal{U}$: szókincs (vagy gondolat, vagy bármi más)
Bemenet: $u \in \mathcal{U}$.
Kimenet: $\mathrm{SF}(u) \in \mathrm{Gen}(u)$.
Adott u-hoz nyelvtípusonként más $\mathrm{SF}(u)$.
Hogyan határozzuk meg $\mathrm{SF}(u)$-t?

Formalizáljuk az OT-t

- Adott $\mathcal{U}$. Bemenet: $u \in \mathcal{U}$.
- Adott Gen. Kimenet: $\mathrm{SF}(u) \in \mathrm{Gen}(u)$.

- Adott függvények (constraint-ek) véges halmaza:
  $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$.
  Az egyszerűség kedvéért t.f.h. $\forall i \in \{1, \ldots n\}$-re $C_i : \mathrm{Gen}(u) \to \mathbb{N}_0$.
- Nyelvtan: rendezés $\mathcal{C}$-n. Ez a rendezés nyelvtípusonként változik.
  T.f.h. $C_1 \gg C_2 \gg \ldots \gg C_n$.
- $\mathrm{SF}(u)$ legyen $\mathrm{Gen}(u)$ optimális eleme az adott nyelvtan szerint.

Hogyan értsük ezt?

Formalizáljuk az OT-t

- Adott $\mathcal{U}$. Bemenet: $u \in \mathcal{U}$.
- Adott Gen. Kimenet: $\mathrm{SF}(u) \in \mathrm{Gen}(u)$.

- Adott függvények (constraint-ek) véges halmaza:
  $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$.
  Az egyszerűség kedvéért t.f.h. $\forall i \in \{1, \ldots n\}$-re $C_i : \mathrm{Gen}(u) \to \mathbb{N}_0$.
- Nyelvtan: rendezés $\mathcal{C}$-n. Ez a rendezés nyelvtípusonként változik.
  T.f.h. $C_1 \gg C_2 \gg \ldots \gg C_n$.
- $\mathrm{SF}(u)$ legyen $\mathrm{Gen}(u)$ optimális eleme az adott nyelvtan szerint.

Hogyan értsük ezt?

OT: a tipológia egy lehetséges modellje (emlékeztető)

$\mathrm{Gen}(\sigma\sigma\sigma\sigma) = \{$[suuu], [usuu], [uusu], [uuus]$\}$.

/σσσσ/	EARLY	LATE	NONFINAL
☞ [s u u u]	0	3	0
[u s u u]	1!	2	0
[u u s u]	2!	1	0
[u u u s]	3!	0	1

$\mathrm{SF}(\sigma\sigma\sigma\sigma) =$[suuu]

Formalizáljuk az OT-t (példa)

$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c\in\mathrm{Gen}(u)}\, H(c)$$

- $u = \sigma\sigma\sigma\sigma$ és $\mathrm{Gen}(\sigma\sigma\sigma\sigma) = \{$[suuu], [usuu], [uusu], [uuus]$\}$.
- Magyar nyelvtan: $C_1 = \textsc{Early} \gg C_2 = \textsc{Late} \gg C_3 = \textsc{NonFinal}$.

- Vegyük észre, hogy a táblázat minden sora egy vektor:
  $H([\text{suuu}]) = (0,3,0)$, $H([\text{usuu}]) = (1,2,0)$,
  $H([\text{uusu}]) = (2,1,0)$, $H([\text{uuus}]) = (3,0,1)$.

- Vektorok **lexikografikus** rendezése: $\mathbf{a} \prec \mathbf{b}$ a.cs.a.
  ha létezik i, hogy (1) minden $j < i$-re $a_j = b_j$, és (2) $a_i < b_i$.

Formalizáljuk az OT-t

$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c \in \mathrm{Gen}(u)}\; H(c)$$

Optimalitáselmélet: $H(c) = \big(C_1(c), C_2(c), \ldots, C_n(c)\big)$
nyelvtan: a $C_1 \gg C_2 \gg \ldots \gg C_n$ rendezés.
opt: lexikografikus rendezés $\mathbb{R}^n$-n.

Harmónianyelvtan: $H(c) = \sum^n{}_{i=1}\; w_i \cdot C_i(c)$
nyelvtan: a C_i-khez tartozó w_i súlyok.
opt: min az $\mathbb{R}$-n értelmezett $<$ reláció szerint.

(Elvek és paraméterek: $H(c) = \bigwedge{}_{i=1}{}^n(w_i \vee C_i(c))$
nyelvtan: a C_i-khez tartozó w_i kapcsolók.
opt: a hamis „optimálisabb” mint az igaz.)

Áttekintés

Formalizáltuk az OT-t

$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c \in \mathrm{Gen}(u)}\; H(c)$$

Felmerülő komputációs kérdések:

1. **Generálás:** adott u és $\mathrm{Gen}(u)$. Adott $(\mathcal{C}, \gg)$.
   Vagyis adott $H(c)$ minden $c \in \mathrm{Gen}(u)$-ra.
   Hogyan találjuk meg $Gen(u)$ optimális elemét $(\mathcal{C}, \gg)$ szerint?

2. **Tanulás:** adott u és $\mathrm{Gen}(u)$. Adott $\mathcal{C}$, de rendezés nélkül.
   Adott egy sor $\left(u_r, \mathrm{SF}(u_r)\right)$ tanulóadat.
   Hogyan találjuk meg azt a rendezést (nyelvtant),
   amely a tanulóadatokat reprodukálni tudja?

Formalizáltuk az OT-t

$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c \in \mathrm{Gen}(u)}\; H(c)$$

Felmerülő komputációs kérdések:

1. **Generálás:** adott u és $\mathrm{Gen}(u)$. Adott $(\mathcal{C}, \gg)$.
   Vagyis adott $H(c)$ minden $c \in \mathrm{Gen}(u)$-ra.
   Hogyan találjuk meg $Gen(u)$ optimális elemét $(\mathcal{C}, \gg)$ szerint?

2. **Tanulás:** adott u és $\mathrm{Gen}(u)$. Adott $\mathcal{C}$, de rendezés nélkül.
   Adott egy sor $\left(u_r, \mathrm{SF}(u_r)\right)$ tanulóadat.
   Hogyan találjuk meg azt a rendezést (nyelvtant),
   amely a tanulóadatokat reprodukálni tudja?

Formalizáltuk az OT-t

$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c \in \mathrm{Gen}(u)}\; H(c)$$

- **Generálás:** adott u és $\mathrm{Gen}(u)$. Adott $(\mathcal{C}, \gg)$.
  Vagyis adott $H(c)$ minden $c \in \mathrm{Gen}(u)$-ra.
  Hogyan találjuk meg *Gen*(u) optimális elemét $(\mathcal{C}, \gg)$ szerint?

- Nézzük végig *Gen*(u) minden elemét!
- És ha *Gen*(u) végtelen (vagy praktikus szempontból túl nagy) halmaz?
- Több megoldást is javasoltak már.
- Javaslatom: vegyük észre, hogy az emberi agy is gyakran hibázik!

Formalizáltuk az OT-t

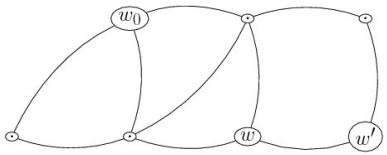
$$\mathrm{SF}(u) \;=\; \arg\,\mathrm{opt}_{c \in \mathrm{Gen}(u)}\; H(c)$$

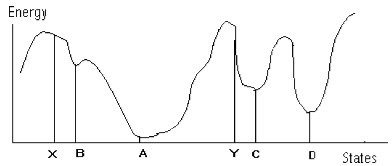
- **Generálás:** adott u és $\mathrm{Gen}(u)$. Adott $(\mathcal{C}, \gg)$.
  Vagyis adott $H(c)$ minden $c \in \mathrm{Gen}(u)$-ra.
  Hogyan találjuk meg *Gen*(u) optimális elemét $(\mathcal{C}, \gg)$ szerint?

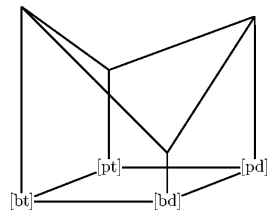
- Nézzük végig *Gen*(u) minden elemét!
- És ha *Gen*(u) végtelen (vagy praktikus szempontból túl nagy) halmaz?
- Több megoldást is javasoltak már.
- Javaslatom: vegyük észre, hogy az emberi agy is gyakran hibázik!

Szimulált hőkezelés (szimulált lehűtés)

(Eddig matematika és informatika. Most statisztikus fizika.)







- Véletlen bolyongás a jelöltek halmazán: szomszédról szomszédra.
- Épp w_1-en állok. Átlépjek-e w_2-re? Átlépés valószínűsége (Boltzmann-eloszlás T „hőmérséklet” mellett):

$$P(w_1 \to w_2 | T) = \begin{cases} 1 & \text{ha } w_2 \preceq w_1 \\ e^{-\frac{H(w_2)-H(w_1)}{T}} & \text{egyébként} \end{cases}$$

- *Heurisztikus/közelítő optimalizáció:* ha elég hosszan bolyongok, jó valószínűséggel megtalálom a (globális) minimumot. De néha lokális minimumba beragadhatok. A sebesség ára a hibázás!

Szimulált hőkezelés (szimulált lehűtés)

Álljon meg a menet! Ha $H(w)$ vektor, akkor mi a szösz $e^{-\frac{H(w_2)-H(w_1)}{T}}$?

- $H(w_2) - H(w_1)$ vektor.
- Legyen T is vektor, és definiáljuk két vektor hányadosát!

Szimulált hőkezelés (szimulált lehűtés)

Álljon meg a menet! Ha $H(w)$ vektor, akkor mi a szösz $e^{-\frac{H(w_2)-H(w_1)}{T}}$?

- $H(w_2) - H(w_1)$ vektor.
- Legyen T is vektor, és definiáljuk két vektor hányadosát!

Hogyan definiáljuk két vektor hányadosát?

Adott két vektor, **a** és **b** $\in \mathbb{R}^n$.
Legyen **b** pozitív vektor, azaz $(0, 0, \ldots, 0) \prec \mathbf{b}$ a lexikografikus rendezés szerint. Ekkor:

$$\frac{\mathbf{a}}{\mathbf{b}} := \sup\{r \in \mathbb{R} | r\mathbf{b} \prec \mathbf{a}\}$$

Két vektor hányadosa kb. az a legnagyobb valós szám, amellyel megszorozva az osztót, még az osztandónál lexikografikusan nem nagyobb vektort kapunk. Pontosabban, a hányados ezen számok halmazának a felső korlátja.

Hogyan definiáljuk két vektor hányadosát?

$$\frac{\mathbf{a}}{\mathbf{b}} := \sup\{r \in \mathbb{R} \mid r\mathbf{b} \prec \mathbf{a}\}$$

Néhány érdekesség:

- Ha ugyanannyi nullával kezdődnek, akkor az eredmény az első nem nulla komponensek hányadosa:
  $(0,0,9,6,2)/(0,0,3,1,1) = 3$, de $(0,0,9,1,2)/(0,0,3,6,1) = 3$ is.
- Ha az osztó több nullával kezdődik, mint a pozitív osztandó, akkor a hányados $+\infty$: $(0,0,9,1,2)/(0,0,0,9,8) = +\infty$.
- Ugyanebben az esetben, ha az osztandó negatív:
  $(0,0,-9,1,2)/(0,0,0,9,8) = -\infty$.
- Ha az osztó kevesebb nullával kezdődik, mint az osztandó, akkor a hányados mindig 0: $(0,0,9,1,2)/(0,2,0,9,8) = 0$.

Szimulált hőkezelés (szimulált lehűtés)

Álljon meg a menet! Ha $H(w)$ vektor, akkor mi a szösz $e^{-\frac{H(w_2)-H(w_1)}{T}}$?

- $H(w_2) - H(w_1)$ vektor.
- Legyen T is vektor, és definiáljuk két vektor hányadosát!
- Ez a hányados valós szám, vagy $\pm\infty$. Ilyen kitevőre emelhető e, és valós számot kapunk.
- Jól lehet szimulálni megfigyelt adatokat, például hangsúly holland gyorsbeszédben.

Példa: hangsúly holland gyorsbeszédben

fo.to.toe.stel ‘fényképező’	*uit.ge.ve.rij* ‘kiadó’	*stu.die.toe.la.ge* ‘ösztöndíj’	*per.fec.tio.nist* ‘perfekcionista’
susu	ssus	susuu	usus
fó.to.tòe.stel fast: *0.82* slow: *1.00*	*ùit.gè.ve.ríj* fast: *0.65* / **0.67** slow: *0.97* / **0.96**	*stú.die.tòe.la.ge* fast: *0.55* / **0.38** slow: *0.96* / **0.81**	*per.fèc.tio.níst* fast: *0.49* / **0.13** slow: *0.91* / **0.20**
fó.to.toe.stèl fast: *0.18* slow: *0.00*	*ùit.ge.ve.ríj* fast: *0.35* / **0.33** slow: *0.03* / **0.04**	*stú.die.toe.là.ge* fast: *0.45* / **0.62** slow: *0.04* / **0.19**	*pèr.fec.tio.níst* fast: *0.39* / **0.87** slow: *0.07* / **0.80**

Szimulált / **megfigyelt** (Schreuder) gyakoriságok.

Áttekintés

Összefoglalás

Hogyan jön össze a nyelv a matematikával és az informatikával?

- A nyelvészeti modellek matematikai alakra hozhatóak.
- Ezek a modellek algoritmikusan implementálhatóak:
  a nyelv „kiszámítása” az agyban és a nyelvtechnológiában.

És a fizikával?

- A fizika a természet jelenségeit írja le a matematika nyelvén.
- A formális nyelvészet (matematikai nyelvészet) a nyelv jelenségeit írja le a matematika nyelvén.

Összefoglalás

Hogyan jön össze a nyelv a matematikával és az informatikával?

- A nyelvészeti modellek matematikai alakra hozhatóak.
- Ezek a modellek algoritmikusan implementálhatóak:
  a nyelv „kiszámítása” az agyban és a nyelvtechnológiában.

És a fizikával?

- A fizika a természet jelenségeit írja le a matematika nyelvén.
- A formális nyelvészet (matematikai nyelvészet) a nyelv jelenségeit írja le a matematika nyelvén.

Összefoglalás

Hogyan jön össze a nyelv a matematikával és az informatikával?

- A nyelvészeti modellek matematikai alakra hozhatóak.
- Ezek a modellek algoritmikusan implementálhatóak:
  a nyelv „kiszámítása” az agyban és a nyelvtechnológiában.

És a fizikával?

- A fizika a természet jelenségeit írja le a matematika nyelvén.
- A formális nyelvészet (matematikai nyelvészet) a nyelv jelenségeit írja le a matematika nyelvén.

Összefoglalás

Hogyan jön össze a nyelv a matematikával és az informatikával?

- A nyelvészeti modellek matematikai alakra hozhatóak.
- Ezek a modellek algoritmikusan implementálhatóak:
  a nyelv „kiszámítása” az agyban és a nyelvtechnológiában.

És a fizikával?

- A fizika a természet jelenségeit írja le a matematika nyelvén.
- A formális nyelvészet (matematikai nyelvészet) a nyelv jelenségeit írja le a matematika nyelvén.

Kutatási lehetőségek

Érdeklődő középiskolás számos nyelvészeti kutatásba kapcsolódhatnak be:

- A nyelvelsajátítás és a nyelvváltozás modelljei.
- Az OTKit programcsomag továbbfejlesztése.
- Kísérletes adatgyűjtés, majd a megfigyelt jelenség számítógépes nyelvészeti modellezése.
- Stb. stb. stb.

Köszönöm a figyelmet!

Biró Tamás:
biro.tamas@btk.elte.hu
http://birot.web.elte.hu/
http://www.birot.hu/

ot kit

Tools for Optimality Theory
http://www.birot.hu/OTKit/