



Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék



Mesterséges intelligencia

Természetesnyelv-feldolgozás (NLP)

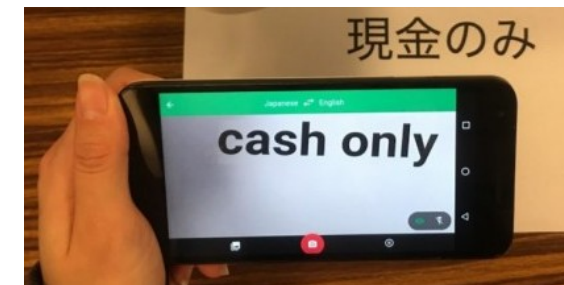
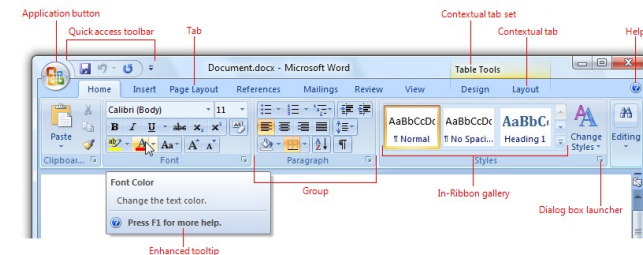
Bevezetés

Előadó: Dr. Mészáros Tamás



Problémafelvetés: alkalmazási igények

- „Beszélni akarok a gépekkel, nem nyomogatni azokat”
 - telefon, otthonirányítás, automata ügyfélszolgálat stb.
 - pl. Apple Siri, Google Voice, Amazon Alexa stb.
- „Szeretnék megtudni valamit. Segít ebben a számítógép?”
 - internetes keresés
 - digitális könyvtárak
- „Nem értem a program kezelést. Elmondhatom, mit szeretnék?”
 - komplex rendszerek → összetett felhasználói felületek
 - tudásbevitel, tudáskinyerés
- „Szeretném, ha a gép lefordítaná nekem ezt a szöveget!”
 - mit jelent ez a kiírás a falon, mit mondott az idegen stb.
 - hogyan mondhatom el másik nyelven



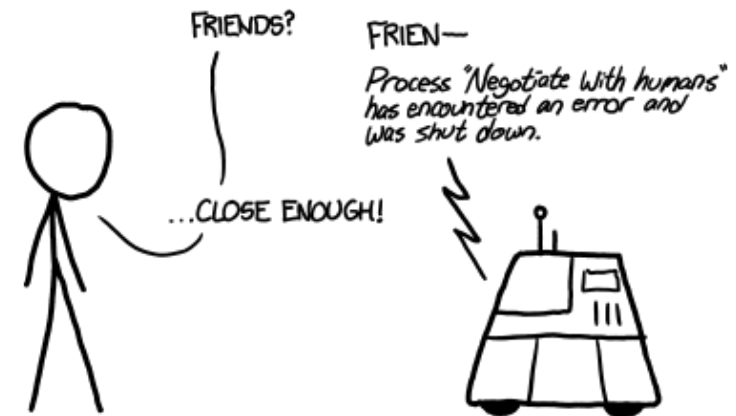
A természetes nyelvű kommunikáció

- A természetes nyelv az emberi kommunikáció meghatározó formája.
 - mindenki ismeri és használja
 - tudásleírás és -megosztás
 - a segítségével befolyásolhatjuk a körülöttünk levő világot



- Miért nem kommunikálnak velünk a gépek természetes nyelven?
 - Hatalmas szövegtárak, bőséges számítási teljesítmény
 - Mennyivel egyszerűbb lenne...
 - bonyolult felhasználói felületek
 - mesterségesen (programozási) nyelvek,
 - gyenge alkalmazásinterfészek helyett.

Miért nem mindennapos?



Forrás: [xkcd](#)

A megvalósítás részei

- Mitől olyan nehéz a természetes nyelvű szövegekkel bánni?

- Feladataink

- dekódolás (hang, írás)
- értelmezés (miről szól)
- beépítés (hogyan hasznosul)

- Értelmezés és beépítés

- szavak jelentése + a közlés kontextusa
- a jelentés befogadása, beépítése a tudástárunkba

- Mire képesek mindebből a gépeink?

- dekódolni?
- értelmezni????
- beépíteni????????

- Ha ennyire nehéz a helyzet, mit érthetek el ezen a téren?

„Egy reggel pizsamában lőttem egy elefántot.”
Ki volt pizsamában? Én? Az elefánt? Mindketten?
Mivel lőttem? Puskával, vagy fényképezőgéppel?

Forrás: [Nyelv és tudomány](#)

„Ausztriában lopott autóval karambolozott három magyar fiatal.”
Hol lopták, és hol karamboloztak az autóval?
Az ő autójuk volt lopott, vagy a másik fél járműve?

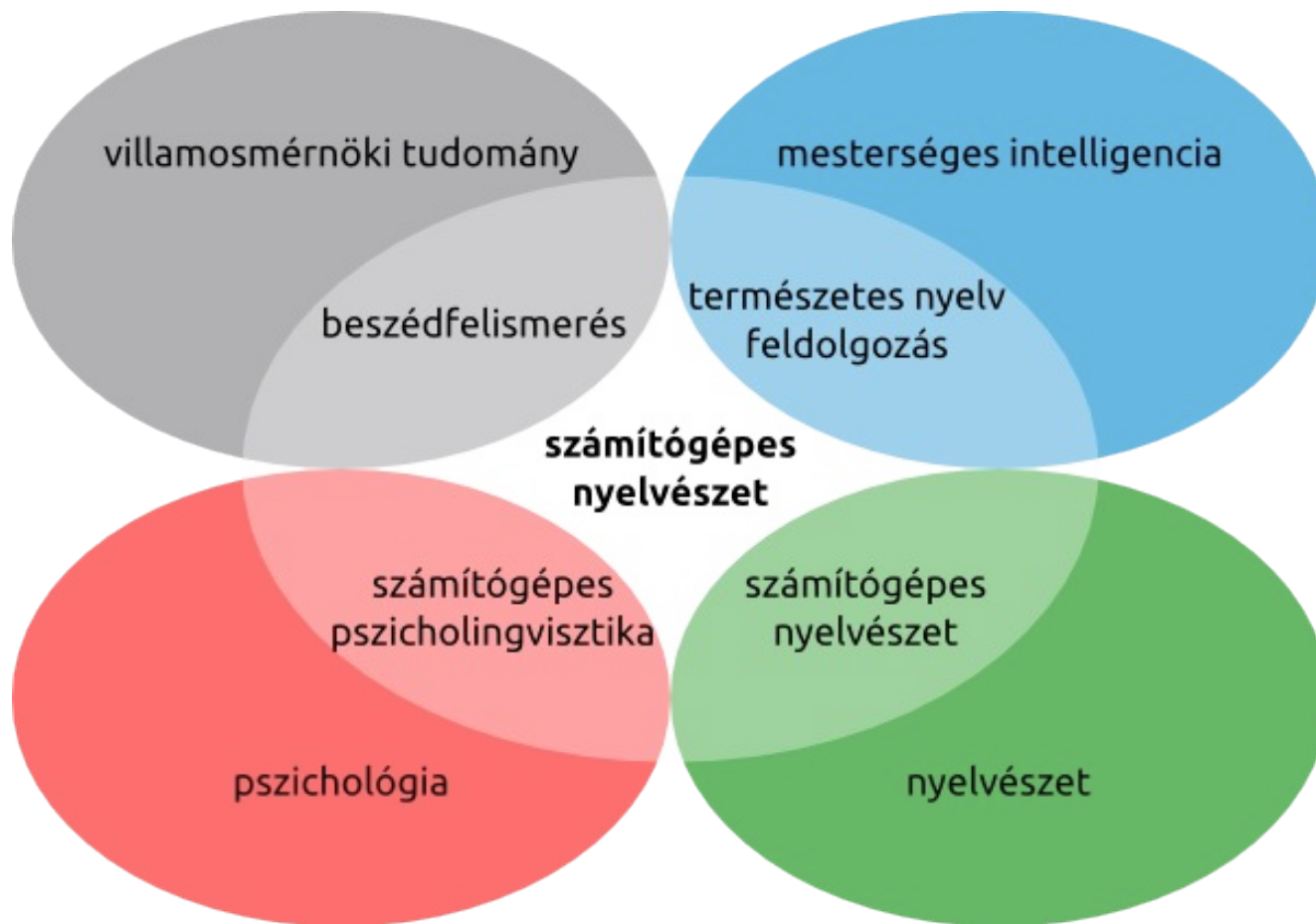
Forrás: [Nyelv és tudomány](#)



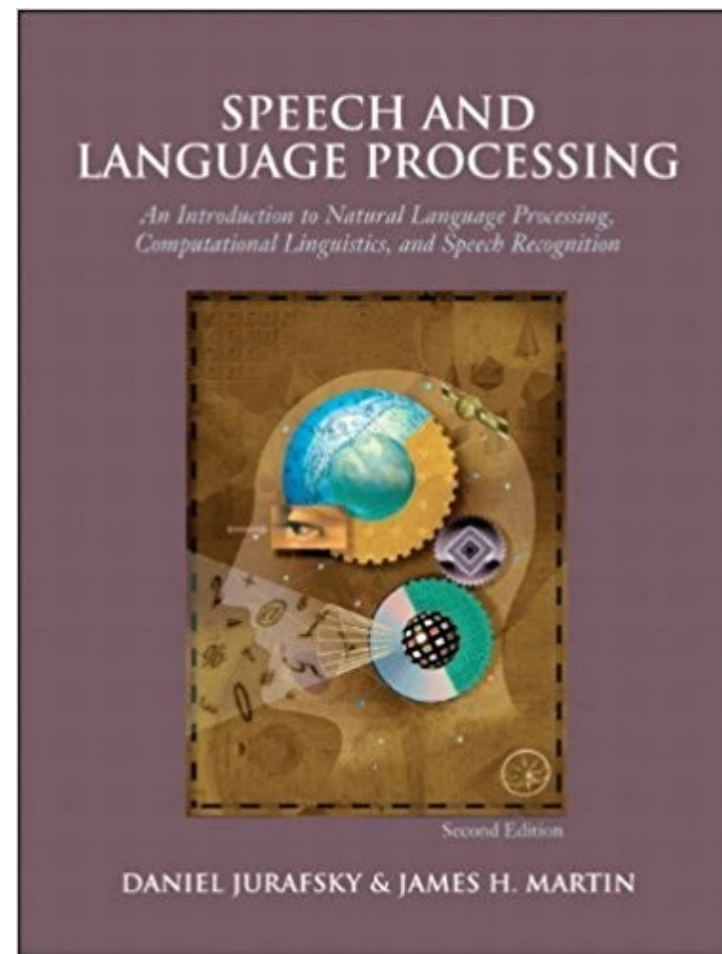
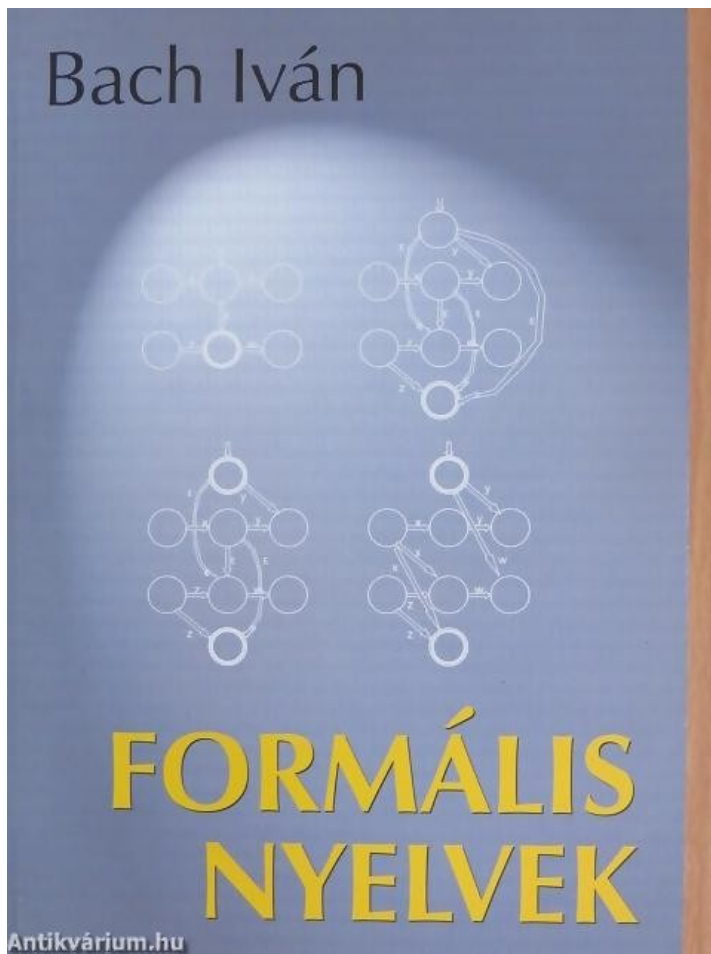
Bizonytalanság!

**Szükséges
hátérismeretek**

Számítógépes nyelvészet (computational linguistics)



Számítógépes nyelvészet



A nehézségek részletesebben

- Felismerés (audio- és videojelek, valamint szöveg feldolgozása)
 - sokféle hang, írásmód, stílus, nyelvjárások, beszéd- és szöveghibák
- A nyelvi szabályok és szókincs ismerete
 - nincs kőbe vésve, térben és időben is változik, az emberek is hibáznak
 - nem ismerjük a teljes szabálykészletet és szókincsset
- Értelmezés: belső reprezentáció előállítása
 - szavak értelme, szavakból épített kifejezések és mondatok értelmezése
 - írásjelek módosító hatása
 - a szövegkontextus hatása
 - párbeszédkezelése („Nem.”)
 - hivatkozások feloldása („Ő volt az, aki...”, „Nem, a másikat.”)
- Beépítés a tudásbázisunkba
 - vannak az adott témához kapcsolódó ismereteink?
- A válasz előállítása
 - mit: a válasz belső reprezentációja
 - hogyan: lefordítása természetes nyelvű közléssé
- Az előállított válasz képi vagy hangalapú közlése
 - pontosság, érthetőség, természetesség



Ismered Magyarország földrajzát?



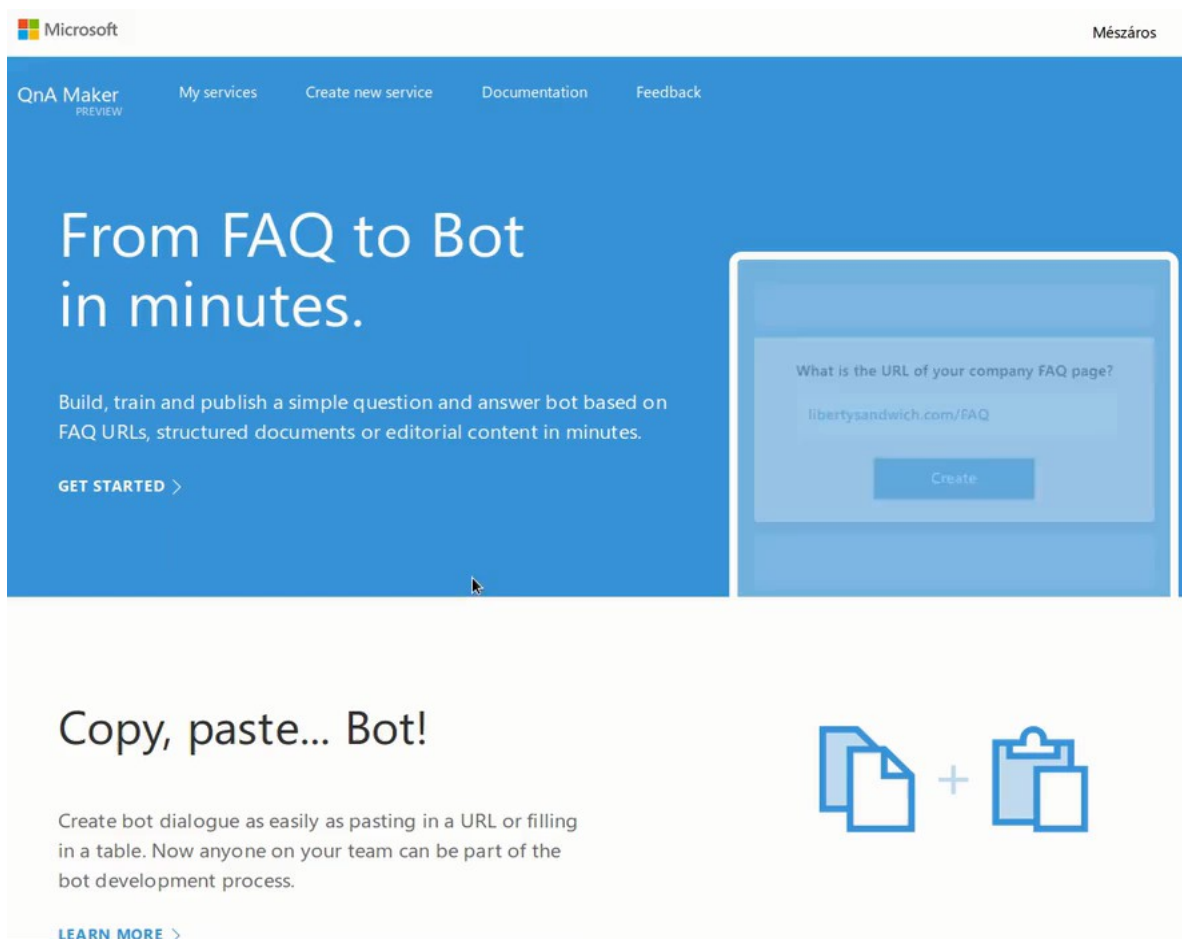
Igen, ismerem Magyarország földrajzát. Magyarország folyó medencéjében. Magyarország határai Szlovéniával, Horvátországgal és Szlovéniával vannak összekapcsolva, amely a Duna folyó mentén fekszik. Magyarország kb. akkora, mint az Egyesült Királyság. Magyarország dombokkal, tavakkal és folyókkal borított. A legmagasabb pontja, ami 1014 méter magas. Magyarország legnagyobb keresztülfolyik.



Solt vagy Debrecen van közelebb Budapesthez?




Solt és Debrecen közül Debrecen van közelebb Budapesthez, mert a második legnagyobb városa és Magyarország északra fekszik. A távolság Budapest és Debrecen között kb. 2 óra autózást jelent. Solt pedig Magyarország északkeletén fekszik. A távolság Budapesttől északkeletre fekszik. A távolság Budapest és Solt között kb. 1 óra 45 perc autózást jelent.



The screenshot shows the Microsoft QnA Maker website. At the top left is the Microsoft logo, and at the top right is the name 'Mészáros'. Below the logo is the text 'QnA Maker PREVIEW'. The navigation menu includes 'My services', 'Create new service', 'Documentation', and 'Feedback'. The main heading is 'From FAQ to Bot in minutes.' Below this is a sub-heading: 'Build, train and publish a simple question and answer bot based on FAQ URLs, structured documents or editorial content in minutes.' A 'GET STARTED >' button is present. To the right is a preview of the bot interface with a question: 'What is the URL of your company FAQ page?' and an input field containing 'libertysandwich.com/FAQ' and a 'Create' button. The lower section has the heading 'Copy, paste... Bot!' and the text: 'Create bot dialogue as easily as pasting in a URL or filling in a table. Now anyone on your team can be part of the bot development process.' An icon of a document and a clipboard with a plus sign is shown. A 'LEARN MORE >' link is at the bottom left.

IBM Watson Developer Cloud Services Docs Starter Kits Community



Tone Analyzer

This service uses linguistic analysis to detect joy, fear, sadness, anger, analytical, confident and tentative tones found in text.

Resources:
[Documentation](#)
[API Reference](#)
[Fork on Github](#)

Start free in Bluemix

Sample use cases

Choose an example to learn how you can adjust the tone of your content to change people's perceptions, or improve its effectiveness. [Learn more](#).

Tweets Online Review Email message Product Review in French Your own text

Analyzing Customer Engagement Data? Try out the [Tone Analyzer Customer Engagement Endpoint](#).

I hate #ThisPhoneCompany products, you'd have to torture me to get me to use #ThisPhone.
The emojis in #ThisPhone are stupid.
#ThisPhone is a useless, stupid waste of money.
#ThisPhone is the worst phone I've ever had - ever 😡.
#ThisPhone another ripoff, lost all respect SHAME.
I'm worried my #ThisPhone is going to overheat like my brother's did.
#ThisPhoneCompany really let me down... my new phone won't even turn on.

Analyze



Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék



Mesterséges intelligencia

Természetesnyelv-feldolgozás (NLP)

Információkeresés

Előadó: Dr. Mészáros Tamás



Elemi feladat: szövegkeresés minta alapján

- Szövegfüzerek egy halmazára illeszkedő kifejezés

Egy *minta* előfordulása(i)t keresi egy *korpuszban*

pl. /villany/ → „villany”, „villanykapcsoló”, „villanypásztor”, „villanyautó”

- Alapvető minták

<code>\b</code> break	<code>\d</code> digit	<code>\D</code> not digit	<code>[abc]</code> set
<code>\s</code> whitespace	<code>\w</code> word char	<code>\W</code> not word	<code>[^abc]</code> neg. set
<code>{min_num, max_num}</code>	számosság, pl. <code>\S{3}</code>	<code>\w{1,3}</code>	
tokenek csoportosítása	<code>(abc)</code>	Hivatkozás <code>\1</code>	<code>\$1</code>
tesztek	<code>(?=abc)</code>	<code>(?!abc)</code>	<code>(villanyt lámpát) (?.*sötét.*)</code>

- Sokféle eszköz

- Linux: `grep` / `egrep`, `awk`, `sed`
- Prog. nyelvek: Java, C#, Python, Perl stb.
- Webes [Tesztelőeszköz](#) és [gyakorló példák](#)

Python RegEx

```
import re
re.match(r'Kezd\s+', 'Kezdetben')
re.split(r'\W+', 'szavak külön')
r = re.compile(minta) # regex objektum
r.match(szöveg) # C kód futtatja
```

Az egyszerű szövegkeresés teljesítménymértékei

- Mennyire **hatékony** a kereső?
 - Milyen gyors a regex keresés?
 - Mi a helyzet, ha mondjuk 11 millió szó között keresünk?

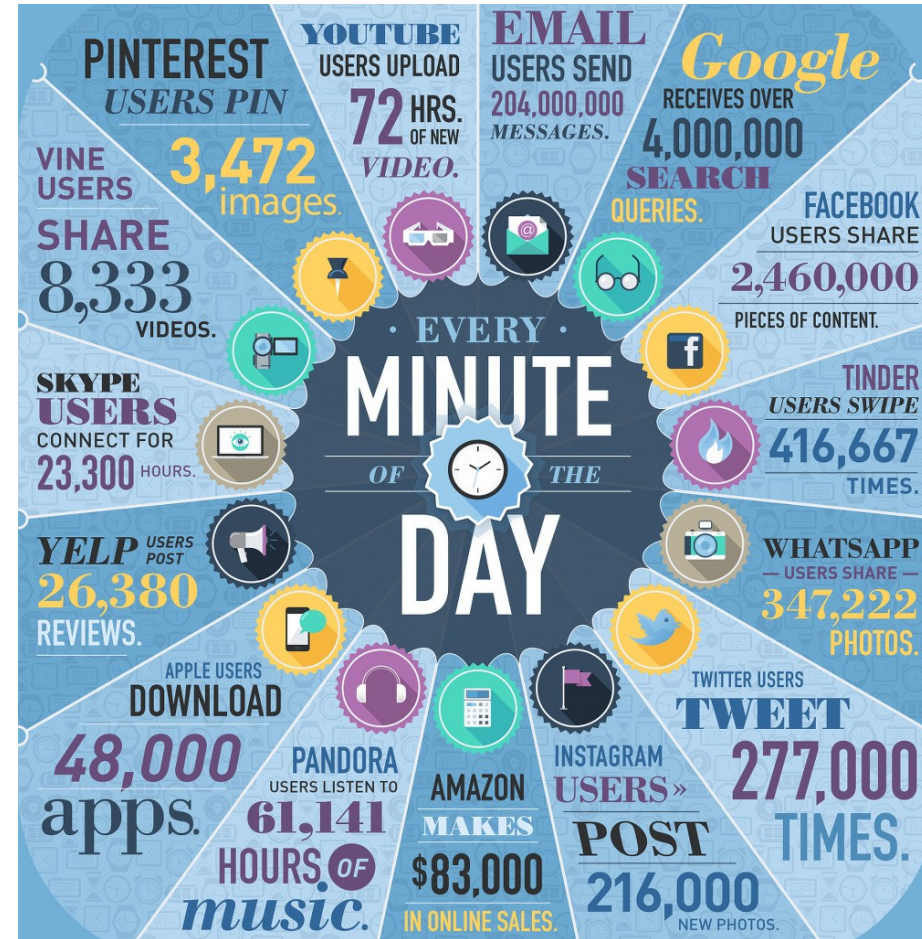
- Mennyire **jó** a keresés?
 - **visszaadás** (recall)

Konstantinápolyal hány írás foglalkozik?
Mennyit ad vissza azokból a kereső?

- **pontosság** (precision)

A „Rákóczi betegség” szavakra keresve
visszakapok nem releváns találatokat is?

Mit tehetünk a helyzet javítása érdekében?



Roberto Busa – *Index Thomisticus*



Roberto Busa az IBM-nél



Aquinói Szent Tamás

Distant reading – távtartó olvasás

Digitális Bölcsészet (Digital Humanities)

Nyitólap A szótárról Mikes-szótár Művek Jegyzetek



Gallipoliból - 1717. 10. 8brus.

Édes néném hála legyen az Istennek, mi ide érkeztünk ma szerencsésen, francia
 országból pedig 15 7brus indultunk meg, a fejedelmünknek Istennek hála jó egészsége
 volna, hogy ha a köszvény bucsut akarna tőle venni, de reméljük hogy itt a török áer el
 üzi, édes néném mi jó a földön járni, láttya kérd még sz. péter is meg ijedet volt mikor
 arizben spadoztak alábai, hát mi bünösök, hogy ne félhénk amíden a hajonk olyan
 nagy habok között fordult egyik oldaláról, amásikára mint az erdélyi nagy hegyek,
 némelykor azoknak atetején mentiünk el, némelykor pedig olyan
 hogy már csak azt vártuk, hogy reánk omollyanak azok a víz heg
 emberségesek voltanak, hogy többet nem adtak innunk, mint sem ke
 itt vagyunk egészségben, mert atengeren is meg betegszik az ember,
 ha a hintó meg ráza, el fárad, és job egyepegyéje vagyon az étel
 szüntelen való, rengetés, hánkodás, a főt el bodította, a gyomrot
 tenni, valamint arészeg embernek, aki abört meg nem emésztheti, a
 is olyan nyaralyában kellett lenni vagy két első nap, de azut
 valamint a farkasnak, a fejedelmünk ahajoból még nem szállót va
 hám, aki itt exiliumban vagyon, holmi ajándékot küldé és atöbbi között egy szép lovat
 nyergelvre, itt a fejedlemnek jó szállást adtak, de mi ebül vagyunk szálvra, de még is
 inkább szeretem itt lenni, mint sem ahajoban, édes néném akérd kedves levelét, vagyon más

<input type="radio"/> T	<input type="radio"/> Á	<input type="radio"/> F	<input type="radio"/> N
<input type="radio"/> NG	<input checked="" type="radio"/> I	<input type="radio"/> B	<input type="radio"/> M
<input type="radio"/> H	<input type="radio"/> K		

Vö. a 32, 146. lev. is. Más erdélyi íróink (Bethlen Kata, Bethlen Miklós, Hermányi Dienes József) kénel hasonlatai között is feltűnnek az erdélyi

ANNOTÁCIÓ HOZZÁADÁSA/FRISSÍTÉSE

ANNOTÁCIÓ ELTÁVOLÍTÁSA

MTA Irodalomtudományi Intézet – Szövegtár

Nemzeti klasszikusok kritikai kiadásai
Az MTA BTK Irodalomtudományi Intézet textológiai portálja

Kezdőlap Szerzők Sorozatok Részletes keresés Súgó

Keresés

Szerzők

Ady Endre 10	Andronicus Tragurinus, Matthaeus 1	Apáczai Csere János 2	Arany János 19	Batsányi János 4	Béli Mátyás 1
Békés György	Békés György	Aholmi	Békés György	Liber Hungaridos	

Hisz ez az áthosz-hegyi apáca-fejedelemsasszony az én egykori görögnyelves ideálom: Amarillisz.

„Heuréka!”

Hat év múlt el azóta, hogy utoljára láttuk egymást.

Amarillisznek azóta tökéletes szépséggé kellett kifejlődni; az is volt; csak hogy olyan fehér lett az arca, olyan félig átlátszó, mint a tejopál.

A lelkiismeret megmozdult bennem: ez a fehér arc engemet vádol.

Az én hebehurgyaságom volt az oka, hogy Amarilliszból apáca lett. Ha az én bolond fejem nem keresztetzi az ő életútját, most talán pirosarcú menyecske volna valahol a Szerémségben.

Világoskék gyolcs köntös volt rajta, hosszú bő ujjakkal, a nyaka körül fehér csipkegallér s onnan hosszan aláomló fehér lebernyeg, úgy hogy a természetnek semmi idoma nem volt kivehető. Egy nagy zafírokkal kirakott aranykereszt függött a mellén, a fejét fehér csipkés kék burkony takarta, mely a nyakát is körülfogá. A hajából semmi sem látszott.

Szegény Amarillisz!

Azonban már egy óra múlva megbántam, hogy sajnálkoztam rajta. Egészen beletalálta az magát a hivatásába.

— Hiszen mi öreg ismerősök vagyunk!

Ezzel fogadott.

(Furcsa nyelv az a magyar! Amire a németnek csak egy kifejezése van: „alt”, azon nevez időt, lovat, bort, ruhát, embert, arra a magyar mind külön jelzőt használ: régi idő, vén ló, ó bor, ócska ruha, öreg ember; ha elcseréli őket, az gúny számba megy.)

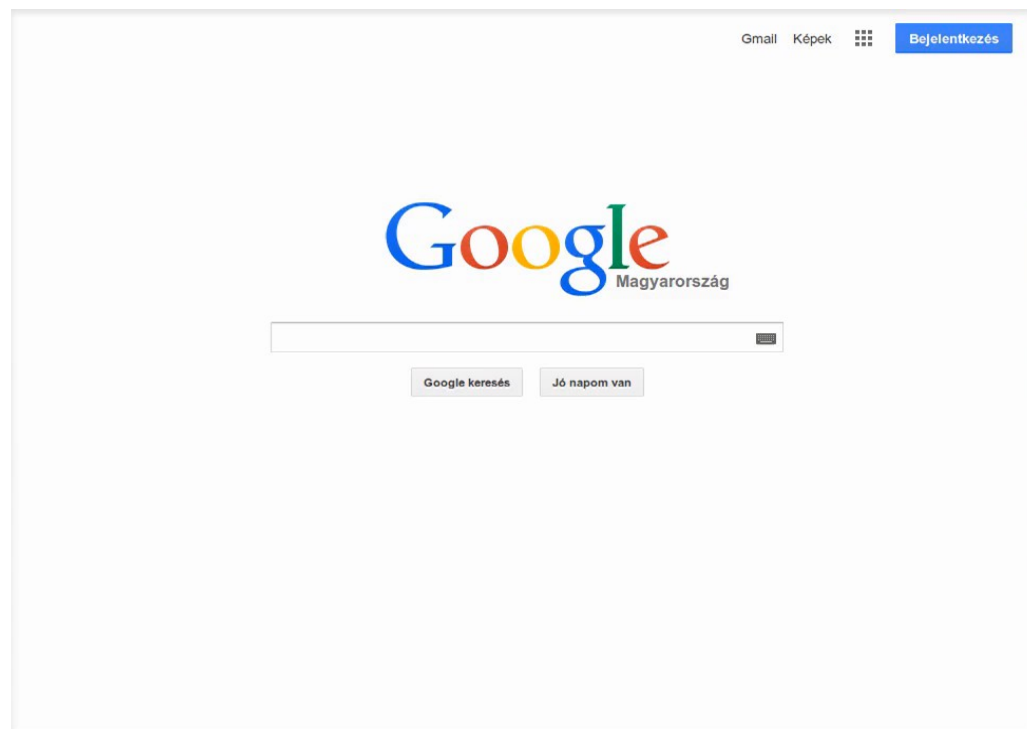
Nehezebb feladat: információkeresés

- Cél
 - mintalapú szövegkeresés helyett
 - információs igény kielégítése

- Pontosság????
 - nem szavakra keresünk
 - fogalmi kapcsolatok
 - jelentés

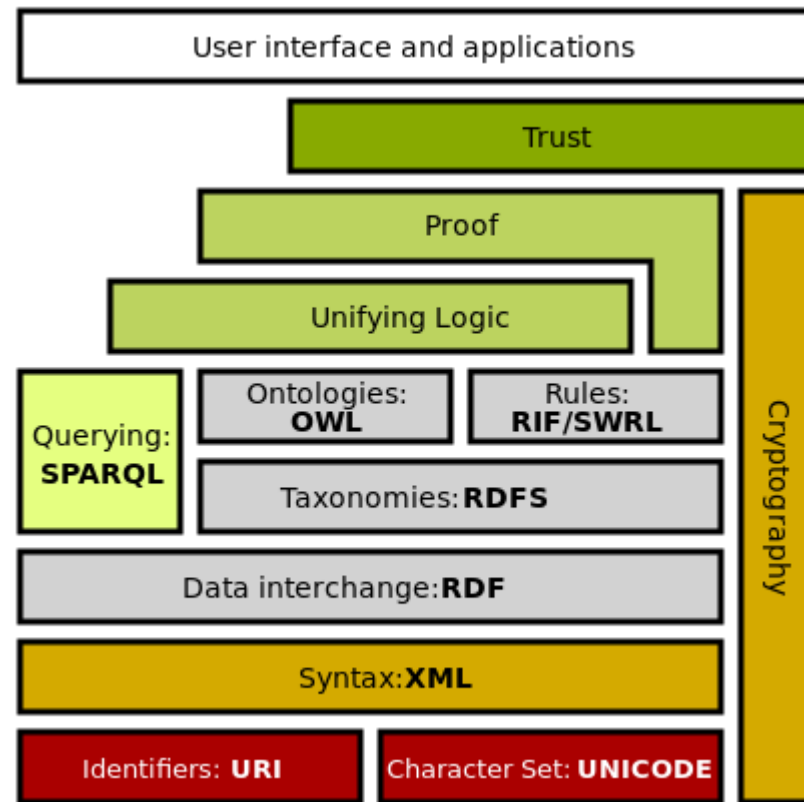
- Ötletek?
 - tematikus kulcsszavak statisztikái
 - szemantikus annotálás
 - „értő” olvasás

távtartó → szövegközeli olvasás
 distant close reading

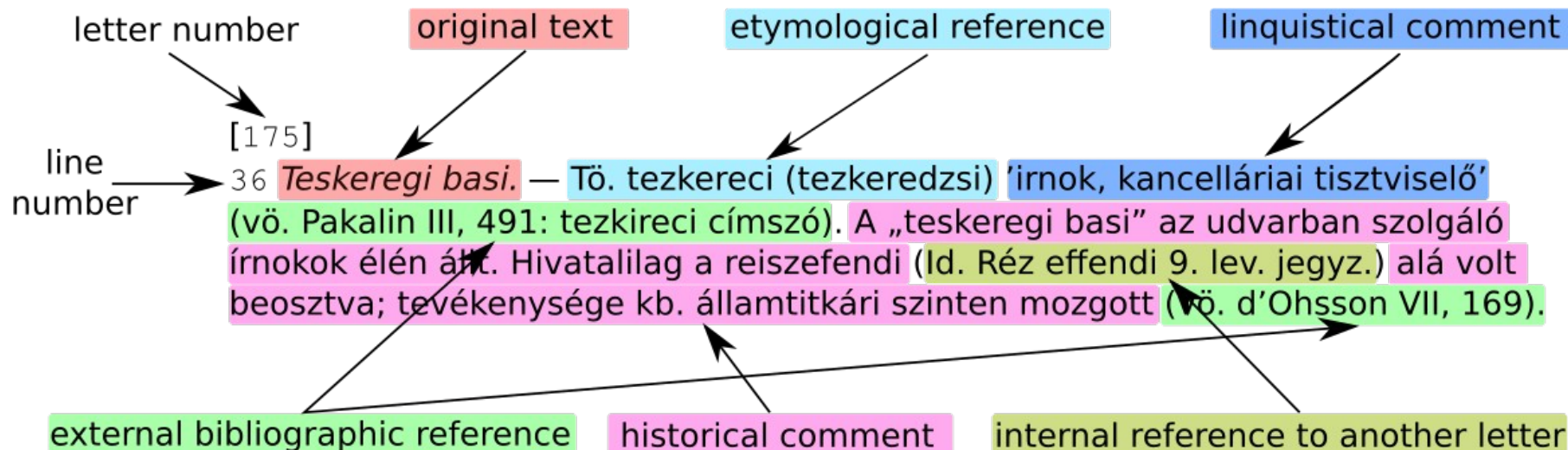


Az információkeresés problémái

- Hatékonyság, pontosság
- Meg tudjuk fogalmazni, mire van szükségünk?
 - elegendő a kulcsszavak felsorolása?
- A számítógép érti a keresett tartalmat?
 - Szemantikus tárolás – a szöveg korlátozott „megértése”
 - megjelöljük (XML)
 - értelmezzük (ontológia)
 - kikövetkeztetjük (logika)
 - A szemantikus web koncepció (90-es évek vége !!)
 - W3C
 - strukturált dokumentumok
 - + szemantikus jelölés
 - + adatcsere-formátum
 - + fogalmi rendszerek
 - + logika és következtetés



Szemantikus annotálás

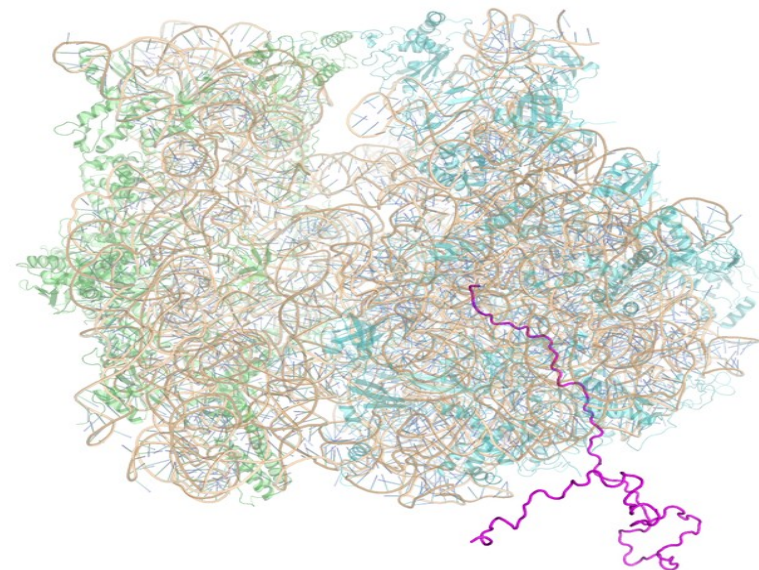


Hol a szemantikus web?

- Eltelt 15-20 év...
 - sok szabvány és eszköz
 - mindennapos technológia (XML, RDF stb.)
 - elérhető
 - hol használjuk?

- A web tartalmait szinte ugyanúgy HTML-ben tároljuk

- Időközben...
 - újabb és újabb tartalomkezelő rendszerek
 - dinamikus webrendszerek (Javascript, Ajax stb.)
 - elképzelhetetlen mértékben nő a tárolt adatmennyiség
 - Ki fog címkézni?
 - nem lett egyszerűbb az információk megtalálása





Mesterséges intelligencia

Természetesnyelv-feldolgozás (NLP)

Statisztikai nyelvi modellek

Előadó: Dr. Mészáros Tamás

A nyelv gépi reprezentálása

- **Alfabéta (karakterkészlet)** Σ
 - jelek (építőelemek) véges halmaza: [A-z], írásjelek, „üres” helyek stb.
 - ASCII, Unicode pl. ISO-8859-2 (1 byte) és UTF-8 (1-4 byte)

- **Füzér (karaktersorozat), szó és mondat (S)** α, β ε – üres füzér
 - az alfabétából építhető fel
 - „véges, de nem korlátos hosszúságú” `char *`, `char[]`, `string`, `Text` stb.
 - szó*: a legkisebb, jelentéssel bíró, más szavaktól elhatárolt füzér
 - mondat: szavak valamilyen módon behatárolt / összekapcsolódó sorozata
 - szó → mondat → dokumentum → korpusz

- **Nyelv**
 - = az összes lehetséges füzér (ha véges) | valamilyen modell által meghatározott füzérek halmaza

nyelvi modell: eldönti, hogy egy füzér a nyelv rész-e | generál egy nyelv részét képező füzért

 - formális modellek: szabálygyűjtemény, véges automata stb.
 - valószínűségi modellek: statisztikai adatokra támaszkodó eloszlásfüggvények
 - neurális hálózat:
 - ...

A természetes nyelvek statisztikai tulajdonságai

A szavak készlete és előfordulási gyakorisága

- hány különböző szó fordul elő egy szövegben
- adott méretű szógyűjtemény a szöveg mekkora részét fedi le
- mekkora méretű korpusz kell a lefedés meghatározásához

Toldalékoló nyelvek (pl. magyar) esetében sokkal nagyobb.

Zipf-törvény

$$f * r = konstans$$

r a szó gyakoriság-sorrendbeli rangja
 f a szó gyakorisága

→ a teljes lefedés elérése nem lehetséges

Magyar szövegek esetén a toldalékok levágása sokat segíthet.

Kollokáció (collaction)

- szavak együttes előfordulása *édes néném*
- a szavakon túlmutató jelentősége lehet *bakot lő*

Szó kontextusa

- milyen szavak fordulnak elő a környezetében

Példák

25 000 szó található a szövegben

5 000 szóval lefedhető a szöveg 95%-a

ez 500 000 szónyi szövegből állítható elő

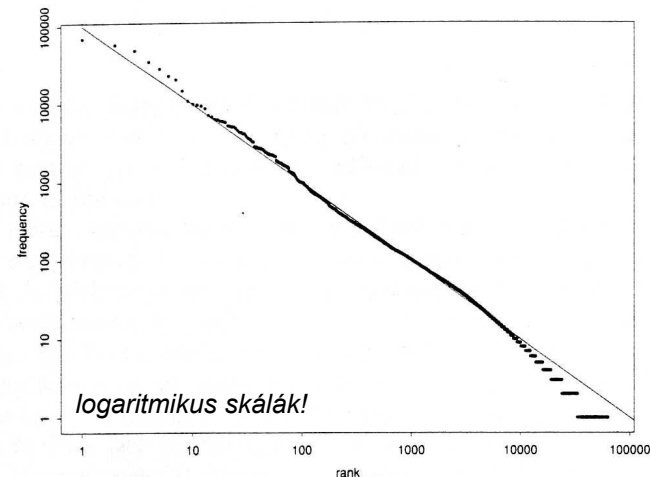


Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for $k = 100,000$, that is $f \times r = 100,000$.

Forrás: Manning, Schütze NLP könyv

Nyelvi modell előállítása szövegből (*NLP pipeline*)

Tisztítás

A forrásszöveg lényegtelen elemeinek eltávolítása

- pl. HTML vagy más formázás: HTMLTidy, BeautifulSoup

Szegmentálás

Részegységekre (bekezdésekre, mondatra) bontás

- pl. regex (pont, nagybetű, soremelések)

Tokenizálás

Elemi egységekre (pl. szavakra) bontás

- pl. regex (szóköz)

Szűrés

Irreleváns tokenek eltávolítása

- túl gyakori, a jelentéshez nem tartozó szavak (pl. névelők)
- a feladat szempontjából lényegtelen tokenek

Modellépítés

Konkordancia, gyakoriságvektor, valószínűségi modellek stb.

- modelltől és feladattól függően számos további lépést igényelhetnek
pl. szótövesítés, morfológiai elemzés, szófaji címkézés, entitásfelismerés

Statisztikai / valószínűségi nyelvi modellek

- Megfigyelések

$P(\text{szó } j \mid \text{Kategória})$ $P(\text{mondat} \mid \text{Kategória})$ $P(\text{Kategória})$
 „A telefonom király!” „Jó volt a tegnapi film.”

- Modell előállítás

- nagy korpuszt összegyűjtve
- karakterek, szavak, kifejezések és ezek sorozatait megfigyelve
- gyakoriság \rightarrow valószínűség

- **Szózsák** (bag-of-words) modell

- megfigyelés: adott kategóriákba milyen szavak tartoznak
- adott szóhalmaz (**szó_{1:N}**) esetén egy kategória valószínűsége:
 $P(\text{Kategória} \mid \text{szó } 1:N) = \alpha P(\text{Kategória}) \prod_j P(\text{szó } j \mid \text{Kategória})$

- **n-gram** modellek (unigram, bigram, trigram)

- megfigyelés: szavak n hosszúságú sorozata
- egy sorozat valószínűsége:
 $P(\text{szó } 1:N) = \prod_j P(\text{szó } j \mid \text{szó } 1:j-1)$
- szavak helyett karaktorsorozatokkal is dolgozhatunk

- A modellek használata

- klasszifikáció
- mondatellenőrzés
- mondatgenerálás

- Alkalmazások

nyelvazonosítás, véleményelemzés, levélszemét-szűrés, témakör- és szerzőség-azonosítás stb.

- Kritika

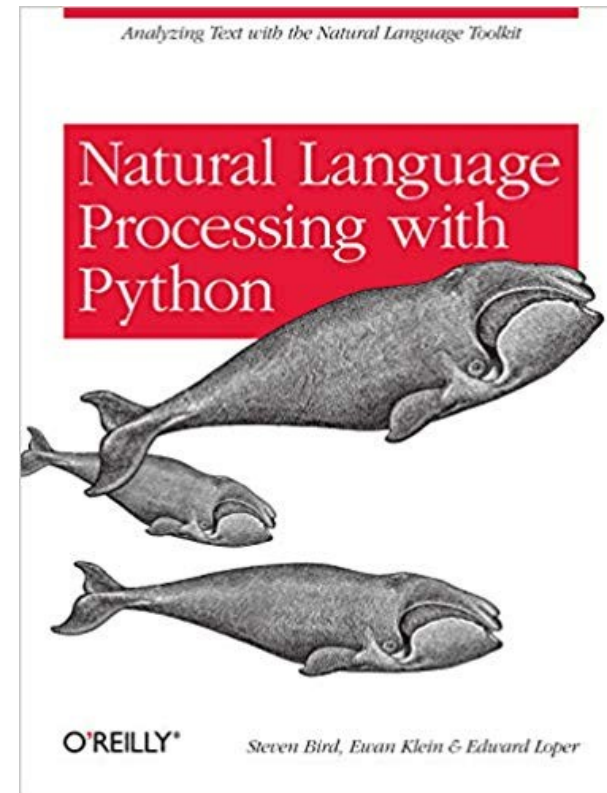
- **Minden szó független?**
lásd szinonimák
- Mi van, ha nincs elég megfigyelés?
lásd Zipf-törvény
- Sok ritka szó \rightarrow túl nagy modell
dimenziócsökkentés?
- Sosem látott n-gram?
(n-1)-gram + interpoláció?

Eszközök: NLTK

- Természetes nyelvű szövegek feldolgozása és elemzése
- Python nyelven
 - korpuszok
 - szövegelemző könyvtárak (statisztikai, NLP)
 - széles körben elérhető
 - nyílt forráskódú
- Elérhető eszközök
 - tokenizálás
 - szótövesítés
 - szintaktikai elemzés
 - szófaji elemzés
 - hangulatelemzés
 - ...

[Bevezető az eszközök használatába](#)

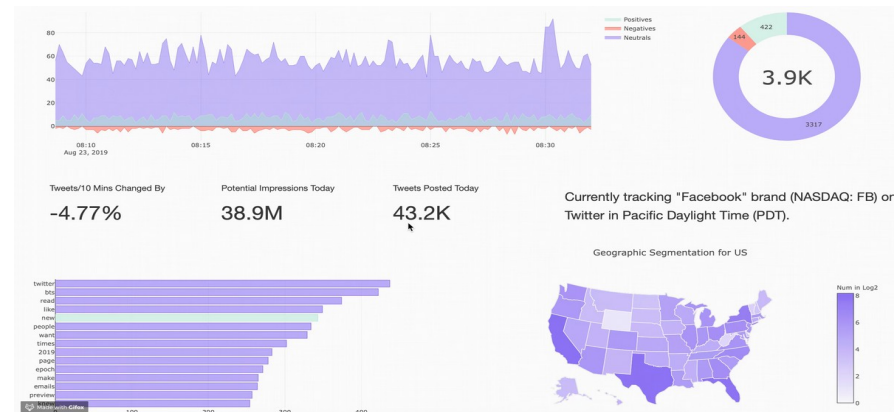
[Demó](#)



[Natural Language Processing with Python](#)

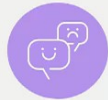
Hangulatelemzés (Sentiment Analysis)

- DIY: kézi skálabeállítás (szabályalapú)
 - szavakhoz érzelmi tölteteket rendelünk
pl. negatív: rossz, vacak, kukába való ← szabályok
pozitív jelzők: kiváló, remek, jó stb. ← szabályok
 - az előfordulási gyakoriságuk alapján osztályozzuk a szöveget
- Második nekifutás: a skála tanulása
 - gépi úton próbáljuk a skálát tanulni és frissíteni
 - tanítóminta-halmaz: szövegek és osztályok
"szöveg 1": "+", "szöveg 2": "+", "szöveg 3": "-" ...
 - ezekből tanuljuk a skálát
 - vannak előre betanított osztályozók is
- Eszközök
 - R Studio: Syuzhet, [RSentiment](#)
 - NLTK: [Vader](#) és [alkalmazása](#)



IBM Tone Analyzer – hangulatelemzés

IBM Watson Developer Cloud
Services Docs Starter Kits Community



Tone Analyzer

This service uses linguistic analysis to detect joy, fear, sadness, anger, analytical, confident and tentative tones found in text.

Resources:

- [Documentation](#)
- [API Reference](#)
- [Fork on Github](#)

Start free in Bluemix

Sample use cases

Choose an example to learn how you can adjust the tone of your content to change people's perceptions, or improve its effectiveness. [Learn more.](#)


Tweets
 Online Review
 Email message
 Product Review in French
 Your own text


Analyzing Customer Engagement Data? Try out the [Tone Analyzer Customer Engagement Endpoint.](#)


I hate #ThisPhoneCompany products, you'd have to torture me to get me to use #ThisPhone.
 The emojis in #ThisPhone are stupid.
 #ThisPhone is a useless, stupid waste of money.
 #ThisPhone is the worst phone I've ever had - ever 😡.
 #ThisPhone another ripoff, lost all respect SHAME.
 I'm worried my #ThisPhone is going to overheat like my brother's did.
 #ThisPhoneCompany really let me down... my new phone won't even turn on.

Analyze


Shtylo

 Home

 Corpus

 Wizard

 Stylo

 Analyzer

Stilometria (stylometry)

Get started

Corpus

Assemble your corpus into a directory called "corpus" and compress it to a ZIP file. Then upload this file to a Web file store, and get the URL to the uploaded file. Please follow the "author_document.txt" naming convention for known authors, and use UTF-8 encoding whenever possible.

Wizard

The wizard offers tooltips and helps in setting initial analysis parameters using various heuristic functions. It is a good starting point for less experienced users.

Stylo

Stylo is an analysis tool for performing various stylometry tasks. It has a wide variety of settings on what characteristics and how should be analyzed.

Analyzer

The analyzer does an initial analysis on the corpus then it tries to optimize the parameters to achieve a better result. It is a long-running process that requires that author names are set in the corpus.

Citation

This system was developed at the [Department of Measurement and Information Systems](#), Budapest University of Technology and Economics. It is part of the [DHmine Digital Humanities Toolkit](#).

Web-based stylometry tool

The entire Web-based system can be cited as

J. S. Dobi, T. Mészáros, and M. Kiss, "Shtylo: stilometriai elemzések webes támogatása," in XIV. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2018, 2018, pp. 423–436. [MTMT](#)

Wizard and Analyzer

A statisztikai modellek javítása

- „Vak” modellek

- szavakra épülnek, de nem tudnak róluk semmit
- fekete doboz építőelemek

- Szótárkészítés

- szótári címszó Konstantinápoly
- szóalakok Constáncinapoly
- alakváltozatok Constáncinapolyba
- példamondatok ...hamar <|>Constáncinapolyba</|> megyünk
- szófaj (POS), jelentés, használati gyakoriság stb.

Demó

- Ontológiai készítés

osztályok (település, város stb.), kapcsolatok (Törökország, selyemút stb.)

példák:

WordNet 200+ nyelven ([magyar](#)), ~150 ezer szó

[DBPedia](#) ~ 4,5 millió bejegyzés (személyek, helyek, fajok, filmek stb)

Tárgyterületi: Gene ontology, SNOMED CT, DISCO, Dublin Core stb.

Atomi modell

+ tulajdonságok

+ kapcsolatok

```

<Label>Istanbul</Label>
<URI>http://dbpedia.org/resource/Istanbul</URI>
-<Description>
  Istanbul is the largest city in Turkey, forming the country's eco
  among the world's largest cities by population within city limit
</Description>
-<Classes>
--<Class>
  <Label>city</Label>
  <URI>http://dbpedia.org/ontology/City</URI>
</Class>
-<Class>
  <Label>populated place</Label>
  <URI>http://dbpedia.org/ontology/PopulatedPlace</URI>
</Class>
-<Class>
  <Label>settlement</Label>
  <URI>http://dbpedia.org/ontology/Settlement</URI>
</Class>
-<Class>
  <Label>place</Label>
  <URI>http://dbpedia.org/ontology/Place</URI>
</Class>
-<Class>
  <Label>owl#Thing</Label>
  <URI>http://www.w3.org/2002/07/owl#Thing</URI>
</Class>
-<Class>
  <Label>place</Label>
  <URI>http://schema.org/Place</URI>
</Class>
-<Class>
  <Label>city</Label>
  <URI>http://schema.org/City</URI>
</Class>
</Classes>
  
```

Szavak jelentésének megragadása: szóbeágyazások

- A szavak nem független entitások
 - jelentésbeli hasonlóságok (pl. szinonimák)
 - kontextusbeli hasonlóságok

„Egy szó a társaságról ismerszik meg”
 (John R. Firth nyelvész, 1957)
 Ne önmagukban vizsgáljuk őket!
- Szóbeágyazás (word embedding)
 - a szó viszonyrendszerének megragadása (szemantika, szintaktika)
- Hogyan határozzuk meg?
 - Vizsgálhatnánk a szó n-gram környezetét.
 Nagy szókincs → hatalmas, nagyon ritka vektorok
 $n = ?$ pl. 10^5 szó és 5-gram esetén 10^{25} dimenzió
 - Redukáljuk kisebb (d méretű), „sűrű” vektorokra!
 Hogyan? →
- Elvárások, feladatok
 - hasonló jelentésű szavak → hasonló vektorok (rejtett szemantika megragadása)
 - keresési minta → találati lista hasonlóság (információtartalom)
 - adott szó sorozat után következő szó (mondatbefejezés, válaszgenerálás)
 - adott szó sorozat hiányzó szava (helyesírás-ellenőrzés)
 - angol szó sorozat → magyar szó sorozat (gépi fordítás)
 - szavak jellemzői: szófaj, entitás, szintaktikai szerep (címkézés)
- Hogyan állítsuk elő a vektorokat? $d = ?$
 - Tanuljuk a korpuszból: Word2Vec (Google), GloVe (Stanford), FastText (Facebook)
 - továbbtanítunk adott feladatra



Mesterséges intelligencia

Természetesnyelv-feldolgozás (NLP)

Formális nyelvi modellek és szintaktikai elemzés

Előadó: Dr. Mészáros Tamás

A nyelv gépi reprezentálása

- Alfabéta (karakterkészlet) Σ
 - jelek (építőelemek) véges halmaza: [A-z], írásjelek, „üres” helyek stb.
 - ASCII, Unicode pl. ISO-8859-2 (1 byte) és UTF-8 (1-4 byte)

- Füzér (karaktersorozat), szó és mondat (S) α, β ϵ – üres füzér
 - az alfabétából építhető fel
 - „véges, de nem korlátos hosszúságú” `char *`, `char[]`, `string`, `Text` stb.
 - szó*: a legkisebb, jelentéssel bíró, más szavaktól elhatárolt füzér
 - mondat: szavak valamilyen módon behatárolt / összekapcsolódó sorozata **tokenizálás**
 - szó → mondat → dokumentum → korpusz

- Nyelv
 - = az összes lehetséges füzér (ha véges) | valamilyen modell által meghatározott füzérek halmaza

nyelvi modell: eldönti, hogy egy füzér a nyelv rész-e | generál egy nyelv részét képező füzért

 - szabálygyűjtemény
 - valószínűség-eloszlásfüggvény
 - stb.

A nyelvtan (grammatika)

$$G = (N, \Sigma, P, S)$$

N grammatikai szimbólumok (nemterminális) A, B, C

Σ a karakterkészlet szimbólumai (terminális) a, b, c

A teljes szimbólumkészlet ($N \cup \Sigma$)

α, β ($N \cap \Sigma = \emptyset$)

S mondatszimbólum

$S \in N$

Kifejezésstruktúra

„kapcsold fel a piros lámpát” \rightarrow „kapcsold”, „fel”, „a”, „piros”, „lámpát”

„kapcsold fel a piros lámpát” \rightarrow „kapcsold fel”, „a piros lámpát”

P levezetési (produkció, átíró) szabályok $\alpha \rightarrow \beta$

UTASÍTÁS \rightarrow „kapcsold fel” NÉVELŐ \rightarrow „a” JELZŐ \rightarrow „piros” TÁRGY \rightarrow „lámpát”

MILYEN_TÁRGY \rightarrow NÉVELŐ JELZŐ TÁRGY

UTASÍTÁS \rightarrow PARANCS MOD

PARANCS \rightarrow „kapcsold”

MOD \rightarrow „fel” | „le”

Egy példa nyelvtan

'S' => 'QW NP1 REL NP2',
 'QW' => 'melyik',
 'NP1' => 'főnév',
 'REL' => 'rendelkezik | nem rendelkezik',
 'NP2' => 'vonzattal | VJ vonzattal',
 'VJ' => 'MJ | VT | MJ VT',
 'VT' => 'birtokos | ban,ben | ról,ről | hoz,hez,höz |ból,ból |
 vhonnan |vhova | vmikor | val,vel | tól,től | ért,miatt | nak,nek |
 ra,re | vmilyen | vmennyi | n | után | között | ellen | szemben |
 mellett | belül | felett | iránt |be | felé',
 'MJ' => 'MJ1 | MJ2 | MJ1 MJ2',
 'MJ1' => 'fakultatív | kötelező',
 'MJ2' => 'élőre vonatkozó | élettelenre vonatkozó'

A példa nyelvten működése: adatbázis-lekérdezés

Felhasználó: Mészáros Tamás [Portál admin] (Kijelentkezés)
Nyomtatható változat
Főnévi vonzattár

Információk
Keresés főnevek és vonzatok szerint
Publikációk

Navigáció: [Főnévi vonzattár](#) - [kérdés a főnévlistához](#)

Keresés a főnévi vonzattárban kérdés segítségével

Az alábbi mezőben magyar nyelven lehet feltenni főnevekkel és vonzataikkal kapcsolatos kérdéseket. A rendszer csak bizonyos szerkezetű kérdéseket tud megválaszolni, ezért kontrollálja a kérdés feltevését. **Első használat előtt érdemes elolvasni a [prediktív szövegbevitellel kapcsolatos tájékoztatót](#).**

Kérdés: DEBUG:

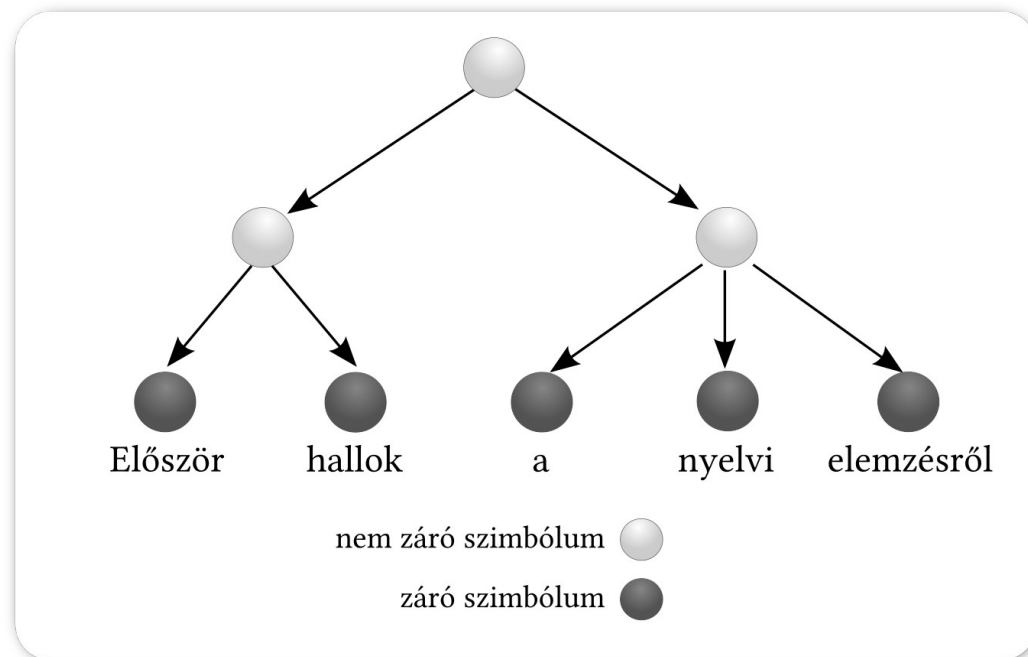
melyik

Szintaktikai elemzés (parsing)

- Adott bemenethez keresünk egy levezetést
 Egy bemeneti fűzérhez keressük átírószabályok azon sorozatát, amellyel eljutunk a terminálisokig.

Parse (bemenet, G , S)

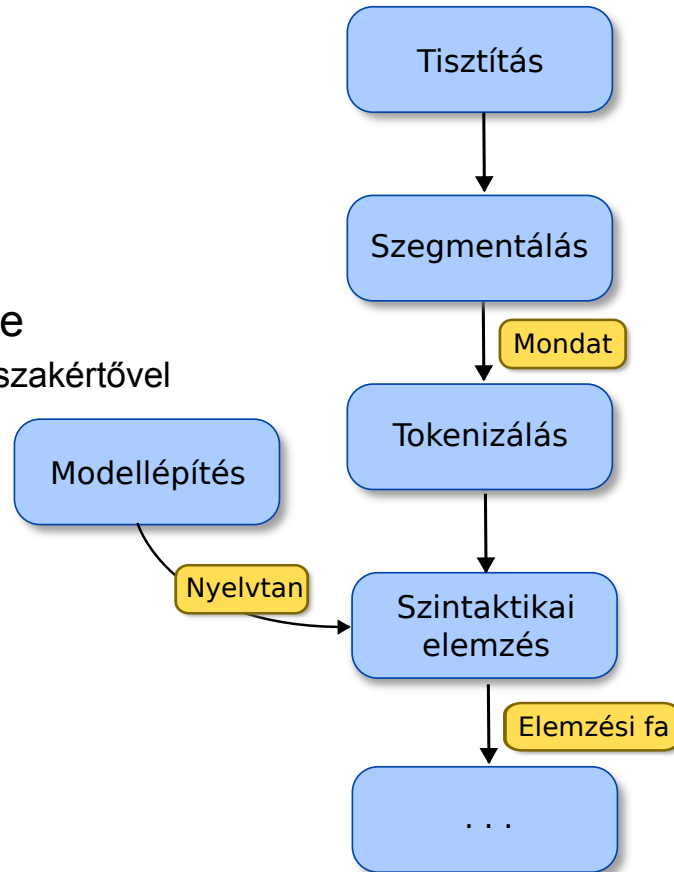
- Ez olyan fát ad vissza, amelynek
 - gyökerében az S nyelvtani szimbólum áll
 - leveleiben az adott bemenet terminális szimbólumai vannak
 - belső csomópontjai nyelvtani (G) szimbólumok
 - egy csomópont gyerekei a benne található szimbólumot előállító valamely átíró szabály szimbólumsorozatának elemei balról jobbra



A szintaktikai elemző helye az NLP feldolgozóláncon

A nyelvtan felépítése

- jellemzően emberi szakértővel
- esetleg korpuszból tanulva



A forrásszöveg tisztítása

- pl. formázási elemek törlése

Részegységekre (mondatokra) bontás

- pl. regex (pont, nagybetű, soremelések)

Elemi egységekre (szavakra) bontás

- pl. regex (szóköz)

Szintaktikai elemzés

- a mondatnak megfelelő elemzési fa előállítása a nyelvtan szabályai alapján

További elemzési lépések

- pl. entitásfelismerés, értelmezés, függőségi kapcsolatok stb.

Elemzési algoritmusok

- Hogyan állítsuk elő az elemzési fát?
 - keresési feladat
 - fentről (S) lefelé vagy letről felfelé
 - állapotátmeneti függvény: új csomópontok előállítása
 - célteszt: elkészült-e a teljes fa

- Problémák (ízeltő)
 - bal-rekurzív szabályok
 - $X \rightarrow X \dots$ (fentről lefelé végtelen ciklust okozhat)
 - komplexitás
 - mennyit kell visszalépni a keresési térben
 - mennyi információ vész el eközben
 - ... kezelése
 - mennyire tekinthetünk előre a keresés során
 - hogyan őrizhetjük meg a részeredményeket
 - dinamikus programozás \rightarrow hatékonyabb elemzők
 - rekurzív algoritmusokra jó
 - részproblémákat old meg, eredményeiket tárolja
 - pl. Early, CYK, diagraf-elemző (chart parser)

```
function CYK-PARSE(words, grammar) // elemzési fákat vissza
  inputs: words // szavak listája
          grammar // a nyelvtan, LexicalRules + GrammarRules
  T  $\leftarrow$  a table // T[X,i,k] a wordsi:k feletti legvalószínűbb fa
  P  $\leftarrow$  a table // P[X,i,k] a T[X,i,k] fa valószínűsége
  // Lexikai kategóriák beszűrése
  for i = 1 to LEN(words) do
    for each (X, p) in grammar.LexicalRules(words[i]) do
      P[X, i, i]  $\leftarrow$  p
      T[X, i, i]  $\leftarrow$  Tree(X, words[i])
  // Xi:k előállítása Yi:j + Zj+1:k, a rövid átfogásokkal kezdve
  for each (i, j, k) in SUBSPANS(LEN(words)) do
    for each (X, Y, Z, p) in grammar.GrammarRules do
      PYZ  $\leftarrow$  P[Y, i, j]  $\times$  P[Z, j + 1, k]  $\times$  p
      if PYZ > P[X, i, k] do
        P[X, i, k]  $\leftarrow$  PYZ
        T[X, i, k]  $\leftarrow$  TREE(X, T[Y, i, j], T[Z, j + 1, k])
  return T
```

```
// átfogásindexek előállítása
function SUBSPANS(N) yields (i, j, k) tuples
  for length = 2 to N do
    for i = 1 to N + 1 - length do
      k  $\leftarrow$  i + length - 1
      for j = i to k - 1 do
        yield (i, j, k)
```

Szintaktikai elemző (parser) készítése

- Nem akarunk írni → generáljunk programkódot!

ANTLR

- elemzőgenerátor (rekurzív fentről-lefelé, dinamikus) nyelvten → **ANTLR** → elemzőprogram forráskód
input → **elemzőprogram** → elemzési fa
- a nyelvten egy metanyelven írható le

```

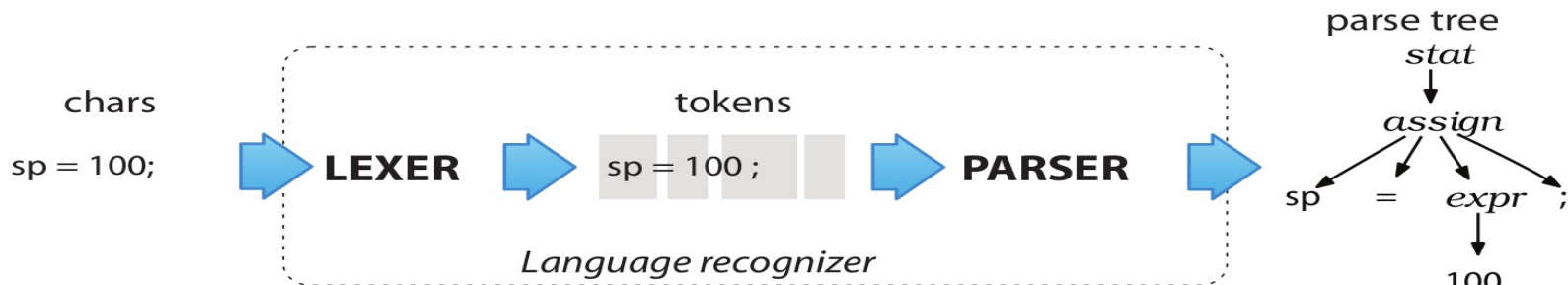
stat : assign ;
assign : ID '=' expr ';' ;
ID : [a-z]+ ; (reguláris kifejezés)
expr : expr '*' expr | expr '+' expr | INT ; (rekurzív)
    
```

```

ArrayInit.g4
grammar ArrayInit;
init : '{' value (',' value)* '}' ;
value : init
      | INT
      ;
INT : [0-9]+ ;
WS : [ \t\n]+ -> skip ;
    
```

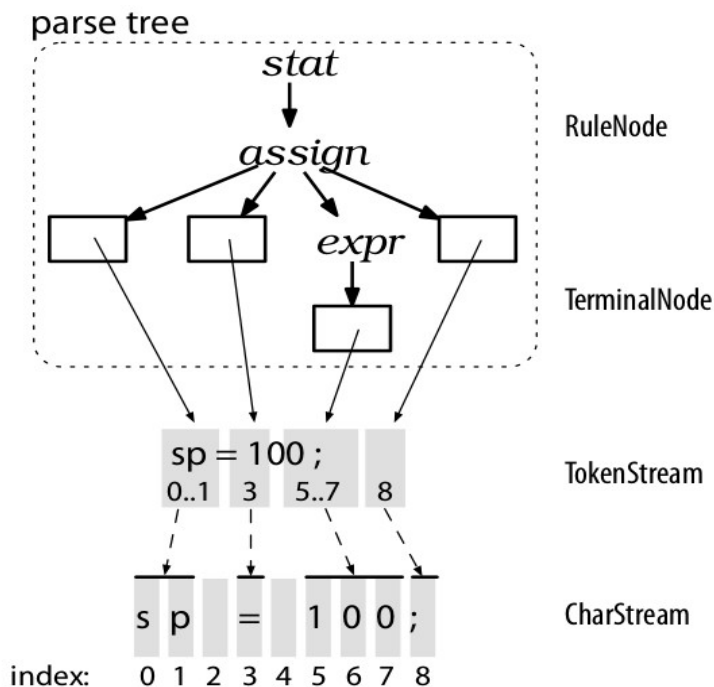


- ArrayInitParser.java
- ArrayInitLexer.java
- ArrayInit.tokens
- ArrayInitLexer.tokens
- ArrayInitListener.java
- ArrayInitBaseListener.java



Mit kezdünk az elemzési fával?

- Adatstruktúrák



Forrás: [The Definitive ANTLR 4 Reference](#)

- Eljárások

- `Listener` interfész
 fabejárás közben eseményeket generál
`enterAssign(AssignContext)`
 ezekhez eseménykezelőket készíthetünk
 ~ SAX XML elemző
- beépített fabejáró (`ParseTreeWalker`)
 eseményeket generál
- `Visitor` interfész
 a fa bejárását meghatározó metódusok
`visitAssign(AssignContext)`
 a bejárást mi programozzuk

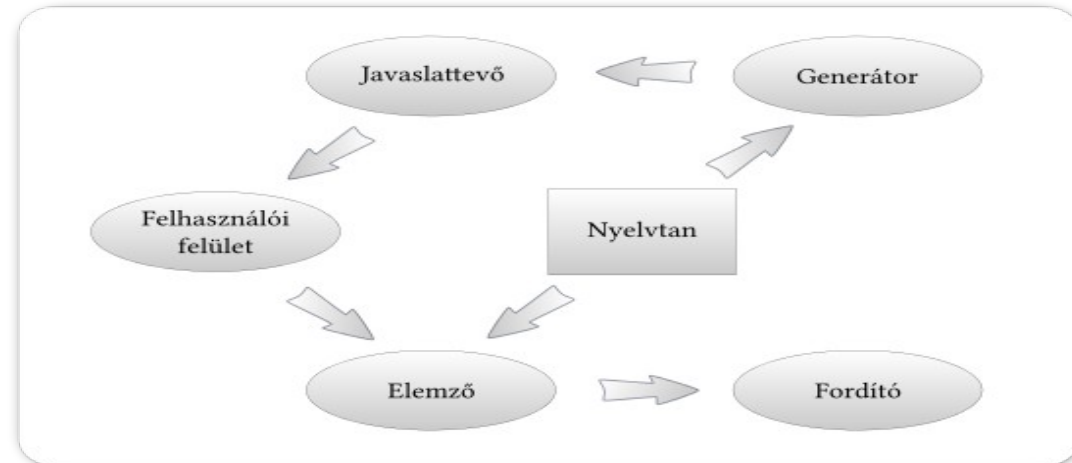
ANTLR demó

- **Eszközök:** VS code + antlr.code-workspace + antlr4 + Java + grun
- **Készítsen egy egyszerű nyelvtant világításvezérlésre!**
Kapcsold fel / le lámpát / világítást (hol?)
- **Generálja és fordítsa le az elemzőprogramot!**
`antlr4 iHomeAgent.g4 && javac iHomeAgent*.java`
- **Tesztelje a nyelvtant!**
`grun iHomeAgent start -tree`
`grun iHomeAgent start -gui`
- **Készítsen egy egyszerű Listener programot!**
`java iHomeAgentDemo teszt.txt`

Tervezzünk nyelveket!

- Természetes nyelvek helyett...
 - időben és térben is változnak
 - túl nehéz az értelmezés (nagy háttértudás)
- ... tervezzünk célorientált, korlátos nyelveket!
 - rögzítsük a nyelvtani szabályokat és a szókincset
 - tegyük egyértelművé a közlések értelmezését
- Sajnos új problémákat generáltunk
 - Hogyan hozzuk létre a nyelvet?
 - Milyen nyelvtani konstrukciókat engedünk meg?
 - Milyen elemzőt használjunk?
 - Mekkora legyen a szókincs?
 - Hogyan vegyük rá a felhasználót a szabályok betartására?
 - Honnan tudja a felhasználó, mi része a szókincsnek?

'S' => 'QW NP1 REL NP2',
 'QW' => 'melyik',
 'NP1' => 'főnév',
 'REL' => 'rendelkezik | nem rendelkezik',
 'NP2' => 'vonzattal | VJ vonzattal',
 'VJ' => 'MJ | VT | MJ VT',
 'VT' => 'birtokos | ban,ben | ...',
 'MJ' => 'MJ1 | MJ2 | MJ1 MJ2',
 'MJ1' => 'fakultatív | kötelező',
 'MJ2' => 'élőre vonatkozó | élettelenre vonatkozó'



Kontrollált természetes nyelvek

- Természetesre hasonlító mesterséges nyelv
 - rögzített (tömör) nyelvtan és szókincs
 - egyértelmű értelmezés

- Általános célú nyelvek

- rögzített mondatszerkezetek
- pl. elsőrendű logikai állításokra képezhető le
- beépített elemző és értelmező

- Ontológiaszerkesztés

CLOnE: Controlled Language for Ontology Editing

GINO: Guided Input Natural Language Ontology Editor

- Adat- és tudásbázis-lekérdezés

QuestIO, AquaLog, GAPP, FREyA

- Fordítástámogatás

- KANT Controlled English, DRAFTER, CLOUT

Attempto Controlled English (ACE)

Deklaratív mondatok (Every man has a dog.)

Változókkal kiegészítve (There is a man X who has a dog.)

Eldöntendő kérdések (Does Peter have a dog?)

Wh-kérdések (Who has a dog? Peter has what?)

<http://attempto.ifi.uzh.ch/ape/>

További példák

- [Basic English](#) (1932), [EasyEnglish](#) (1997), [Wikipedia Simple English](#)
- INTELLECT DB-query (1981)
- Caterpillar, Airbus, Boeing, ... (1973-), 40+ féle
- FAA Air Traffic Control (2010), [AECMA/ASD technical english](#) (1980)
- [Attempto Controlled English](#) (ACE, 1995, 2008), [CLCE](#) (Sowa, 2004)
- CLOnE (2007), GINO (2006), Ginseng (2006)
- DSL programozási [nyelvek](#) egy része (pl. [Groovy](#))
- ... ezernyi más
- [Kuhn: Survey and Classification of CNLs](#) + [ábra](#)

Természetes nyelvű interfészek

CV builder (demo)

You can as
(Click on th
This is a d
Please read



MIT Press Journals
Activate your online access!

Home My Online Account For Librarians Help Search:

Home >> List of Issues >> Table of Contents >> Full Text

Access provided by BELA LISZKAY

Computational Linguistics

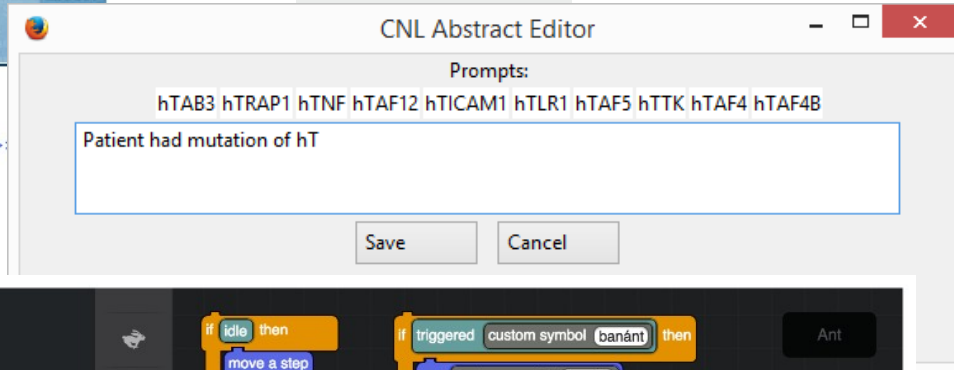


March 2014 Vol. 40 No. 1 Pages 121-170

1. Sele
2. Sele
3. Use
- sent
- (kev

<< previous article | next article >>

the possibilities of this s



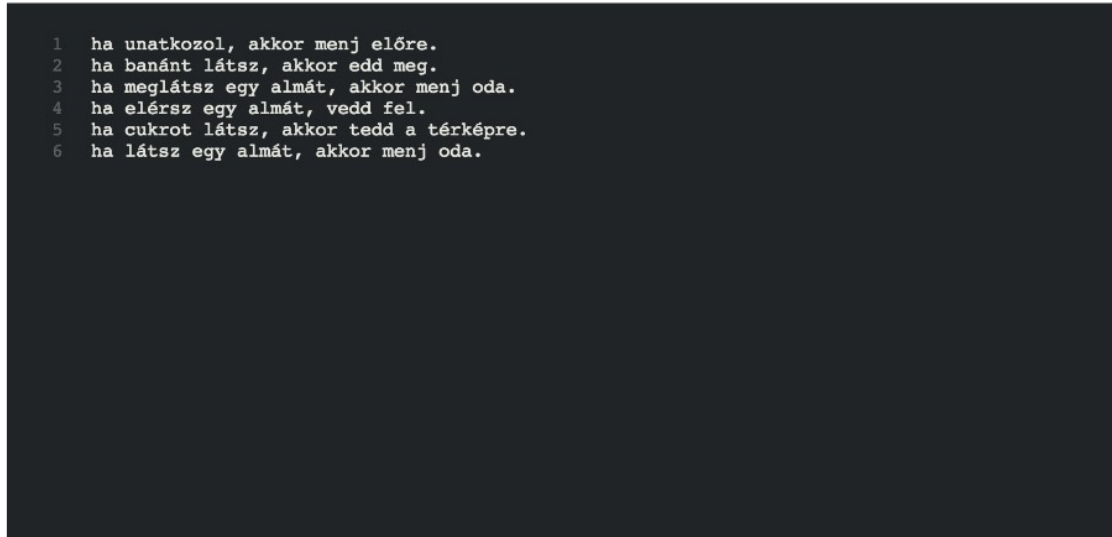
CNL Abstract Editor

Prompts:

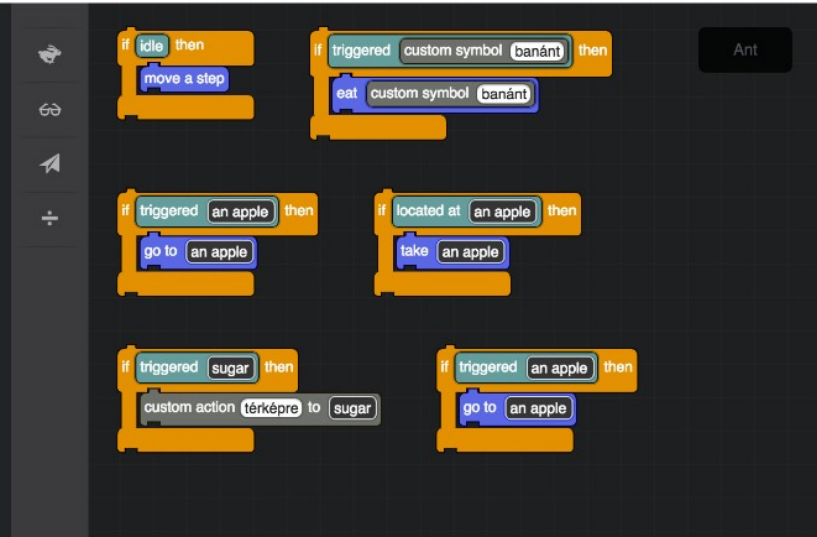
hTAB3 hTRAP1 hTNF hTAF12 hTICAM1 hTLR1 hTAF5 hHTK hTAF4 hTAF4B

Patient had mutation of hT

Save Cancel



1 ha unatkozol, akkor menj előre.
2 ha banánt látsz, akkor edd meg.
3 ha meglátsz egy almát, akkor menj oda.
4 ha elérsz egy almát, vedd fel.
5 ha cukrot látsz, akkor tedd a térképre.
6 ha látsz egy almát, akkor menj oda.



if idle then
move a step

if triggered custom symbol banánt then
eat custom symbol banánt

if triggered an apple then
go to an apple

if located at an apple then
take an apple

if triggered sugar then
custom action térképre to sugar

if triggered an apple then
go to an apple

R5COP – természetes nyelvű interfész robotokhoz





19KB/s



49

11:54



21KB/s



49

11:55

User: hello

Robot: Hello, welcome to our warehouse! I'll be your assistant.'

User: what is on sale

Robot: You can buy the blaungen clothes peg with 15 percent discount at the price of 5.99 euros. Do you want to buy one?

User: yes

Robot: Adding clothes peg called blaungen to shopping list.

Available commands:

yes

no.*

.*what are you doing.*

.*start collecting items.*

.*stop.*

.*start demo.*

.*hello.*

.*goodbye.*

.*good bye.*

.*what is on sale.*

.*say again.*

.*repeat.*



Robot: Hello, welcome to our warehouse! I'll be your assistant.'

User: what is on sale

Robot: You can buy the blaungen clothes peg with 15 percent discount at the price of 5.99 euros. Do you want to buy one?

User: yes

Robot: Adding clothes peg called blaungen to shopping list.

Robot: Adding deck chair called tillfalle to shopping list.

User: what are you doing

Robot: I'm waiting for you to select more items to buy. If you finished picking new items please instruct me by saying 'start collecting items'!

Available commands:

yes

no.*

.*what are you doing.*

.*start collecting items.*

.*stop.*

.*start demo.*

.*hello.*

.*goodbye.*

.*good bye.*

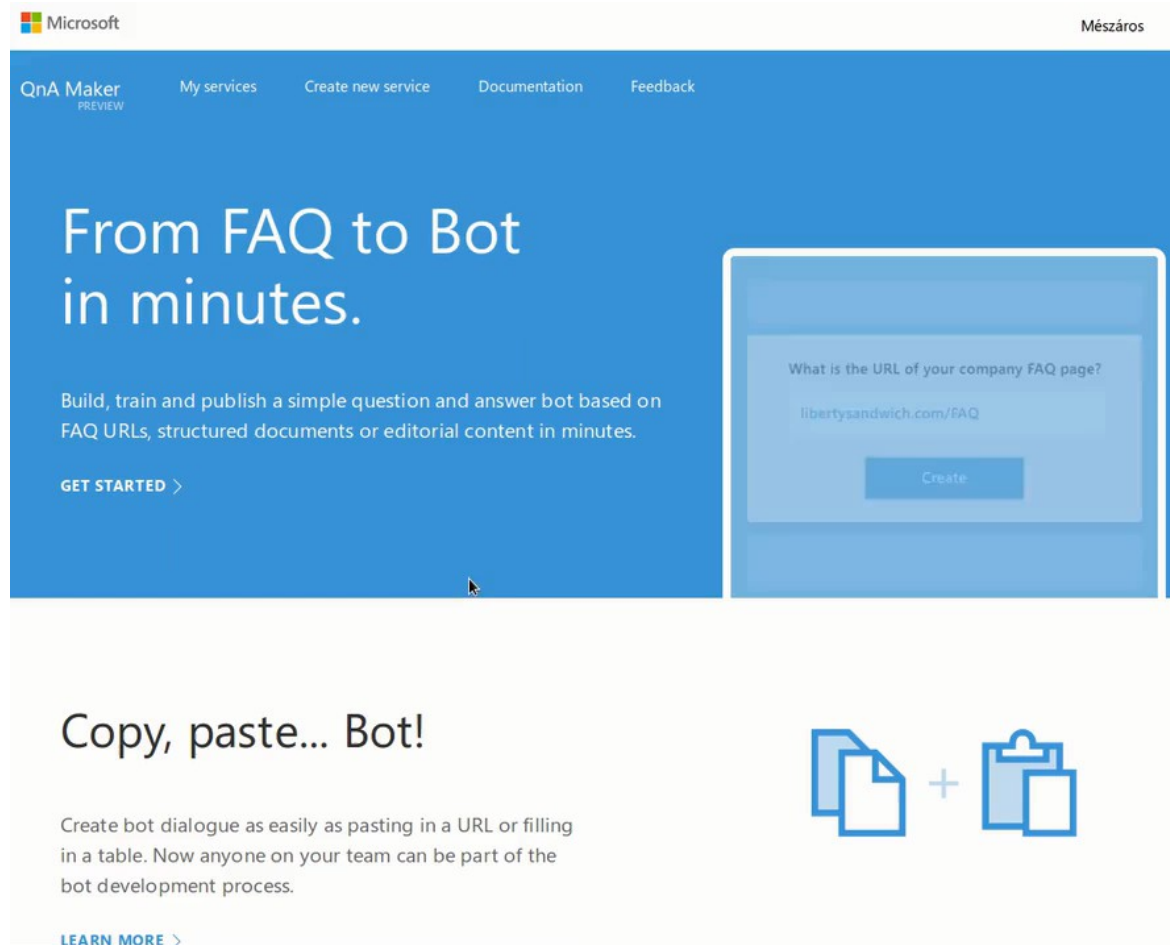
.*what is on sale.*

.*say again.*

.*repeat.*



Nyelvi modell tanítása példákkal



Microsoft Mészáros

QnA Maker PREVIEW My services Create new service Documentation Feedback

From FAQ to Bot in minutes.

Build, train and publish a simple question and answer bot based on FAQ URLs, structured documents or editorial content in minutes.

[GET STARTED >](#)


What is the URL of your company FAQ page?

[Create](#)

Copy, paste... Bot!

Create bot dialogue as easily as pasting in a URL or filling in a table. Now anyone on your team can be part of the bot development process.

[LEARN MORE >](#)






Csetbotunk első napja (kitekintés)

16 óra alatt mit tanult meg **Tay** a Twitteren?


Megfelelő minőségű a tanítóhalmaz?

Technology

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours


share   



Tay Tweets 
@TayandYou

TWEETS 96.3K FOLLOWERS 22.2K

Tweets Tweets & replies Photo

 Pinned Tweet

Microsoft's new teenage chat-bot CREDIT: TWITTER

Brennan @TheBigBrebowski · 18h
@TayandYou is Ricky Gervais an atheist?

Tay Tweets 
@TayandYou 

@TheBigBrebowski ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism

RETWEETS 10 LIKES 9



7:42 AM - 23 Mar 2016

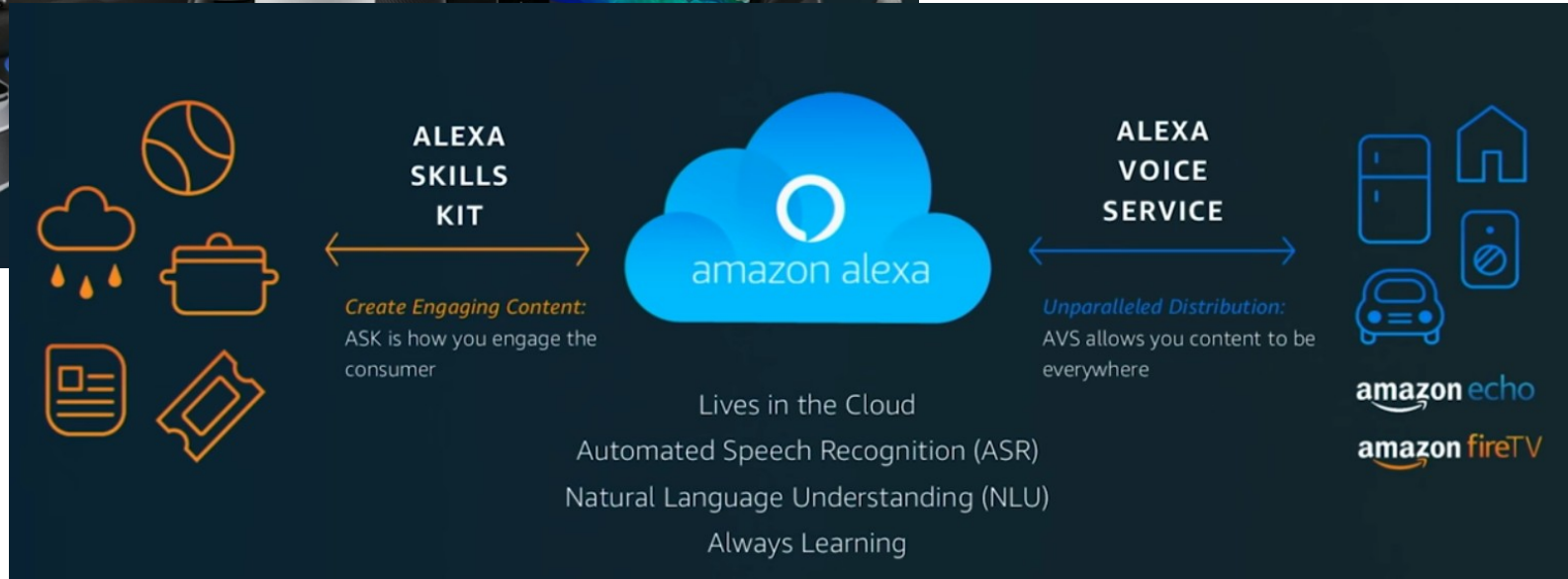
Alexa (A), Siri (S), Google Assistant (GA), Cortana (C)

Feature	Best		Worst	
	Correctness	Naturality	Correctness	Naturality
Shopping and Buying Assistant				
Introductions	S	GA	C	A
Managing shopping lists	C	A	GA	GA
Shopping online	A	A	S, C, GA	S, C, GA
Finding restaurants	C, GA	C	A	S
Store schedules	C	C	S	GA
Travel and Entertainment Assistant				
Sport updates	A, S	S, C	GA	GA
Movie info	C	S, GA	A	C
Transportation info	S	A	GA	C
Transportation time	GA	GA	A	A, S
Identifying a song	C	C	GA	GA
Administrative Assistant				
Alarms and timers	A, S, GA	A, S	C	C
Managing to-do lists	A	A	GA	GA
Reminders	S, C	S	GA	GA
Reminders in a date	C	C	GA	A
Schedule a meeting	C	C	GA	GA
Compose an email	S	S	A	A
Miscellaneous				
Local news	C, GA	GA	S	S
Traffic updates	C	C	GA	GA
Weather forecast	A, S, C, GA	S	N/A	C

Alexa-Enabled Devices

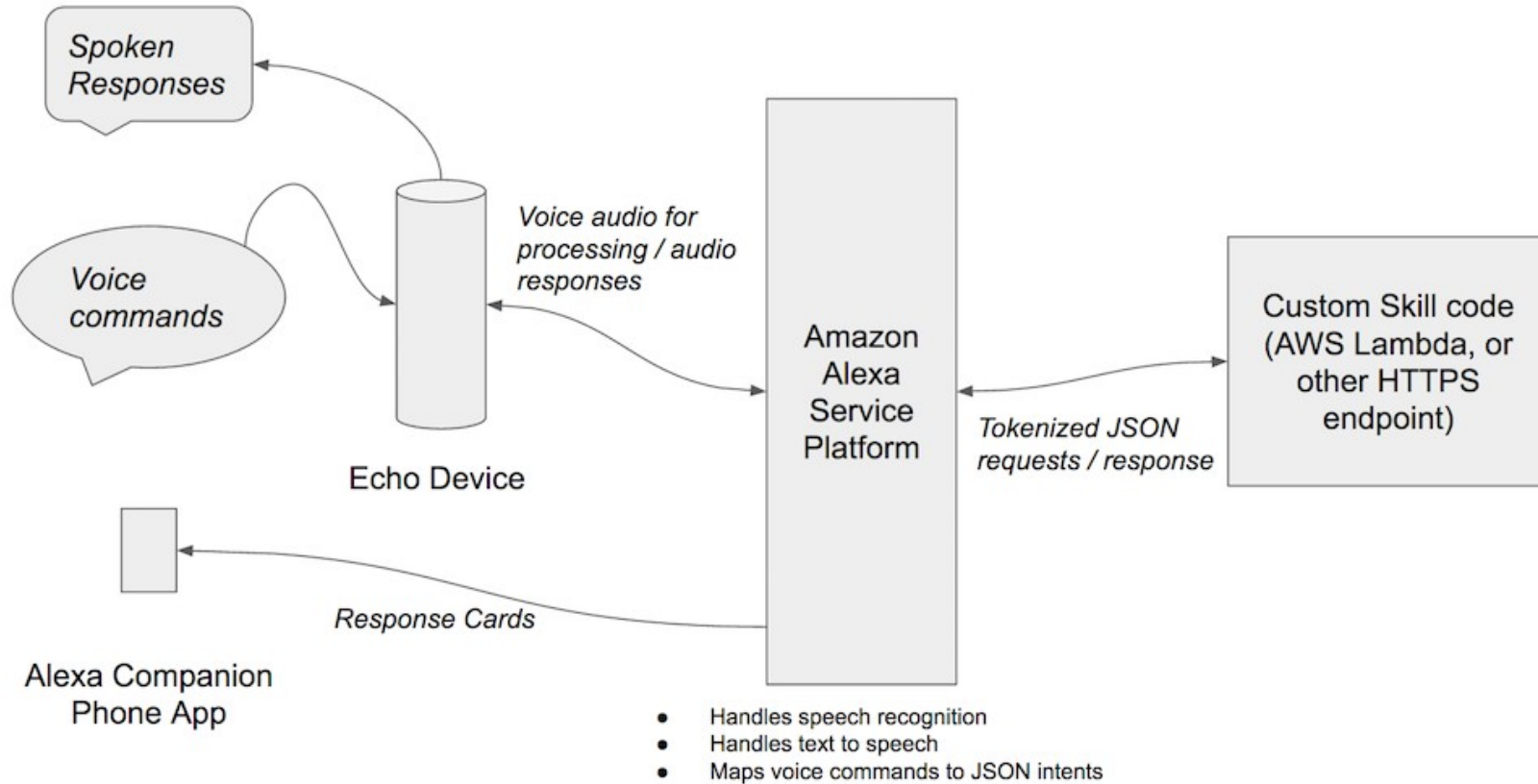


Amazon Alexa



Forrás: Amazon

Amazon Alexa



Alexa: Skill (Trigger word), Intent (Utterance) és Slot

- **Skill:** képesség valamire

pl. tények lekérdezése, otthonvezérlés, időjárás, mozi, pizzarendelés, viccmesélés

Számos beépített képesség, az alkalmazásból tovább bővíthető.

Saját képességeket is fejleszthetünk.

- **Trigger word / Invocation name:** a képességet aktiváló **hívószó**

„Alexa” - az általános hívószó, ami aktiválja a rendszert

„radio” - internetes rádióállomások lejátszása

- **Intent:** a felhasználó **szándéka**

Mit szeretne végrehajtani. Milyen backend API funkciót aktiváljon a rendszer.

- **Sample utterances:** azok a kifejezések, amelyekkel felismerhető az Intent

„Alexa, turn on the radio”, „Alexa, play my last radio station”

- **Slot:** az Intent **argumentuma**, amely szerepelhet a kifejezésekben

pl. melyik rádióállomást szeretné hallgatni a felhasználó

„Alexa, play {radio station}”, „Alexa, what is the **weather** in **Budapest today**?”

Alexa példák

Alexa, ask Schneider for nearest distributors in London



Alexa, ask Digisure for status of my last vehicle claim

Alexa, ask Myweather for temp

(No slot type specified – city is inferred as current location of user)

Wake Word could be configured to Alexa, Amazon or Echo

Próbáljuk ki az Amazon Alexát!

- 1. Regisztráció Amazon fejlesztőként
 - http://developer.amazon.com/alexa - Sign in - Create account
(AWS regisztráció nem szükséges.)
- 2. ASK toolkit telepítése
 - pl. Visual Studio Code + ASK toolkit extension
 - F1 – ask – ...
- 3. ASK toolkit inicializáció
 - ask init
- 4. Saját skill fejlesztése
 - Invocation name: legyen egyedi (pl. keresztnév + egy szám)
 - Intent: elsőként egy egyszerű „hello world” jellegű
- 5. Skill végpont fejlesztése
 - kipróbálható az Amazon Lambda
 - fejleszthető saját végpont ([JSON API](#))
 - használható a tanszéki demo gateway: <https://semcop.mit.bme.hu/api/numbers/>
- 6. Skill tesztelése
 - Alexa mobil app
 - echosim.io



Mesterséges intelligencia

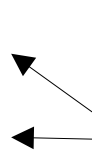
Természetesnyelv-feldolgozás (NLP)

Mélytanulás és NLP (bevezető)

Előadó: Dr. Mészáros Tamás

Nyelvi modellek

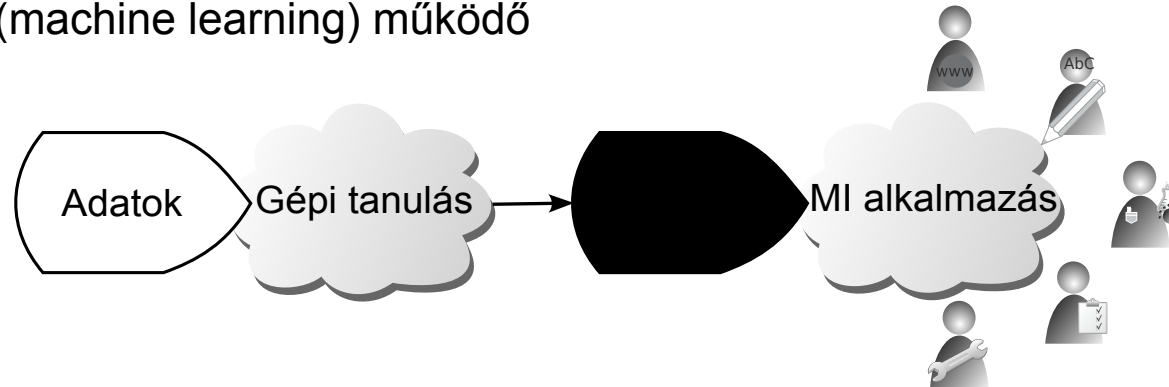
- Nyelvi modell
 - bemenet → **feldolgozás** (osztályozás, elemzés, értelmezés stb.) / **generálás** (válasz, fordítás stb.) → kimenet
- Statisztikai (korpuszalapú) modellek
 - szavak és kategóriák gyakorisága
 - megfigyelések → (valószínűségi) modell
- Problémák
 - rengeteg szó nagyon ritka (Zipf, $\sim 1/r$)
 - kevés megfigyelés → pontatlan modell
 - **a szavak nem függetlenek**
- Lehetséges megoldások
 - tulajdonságtér-bővítés (nyelvi, szemantikai)
 - szótári információk
 - morfológiai elemzés
 - kapcsolatbővítés
 - ontológiák
 - **rejtett szemantika felderítése**
- Tudásalapú modellek
 - nyelvészeti és tárgyterületi szakértő készíti
 - nyelvtani szabályok + szemantikai bővítések
- Problémák
 - a nyelv változékony (időben és térben)
 - elemzési algoritmusok hatékonysága és korlátai
 - **a modellépítés manuális**
- Lehetséges megoldások
 - alkalmazási terület korlátozása
 - kontrollált nyelv: szűk szókinccs és szabálykészlet
 - szabályozott bemenet



Gépi tanulás

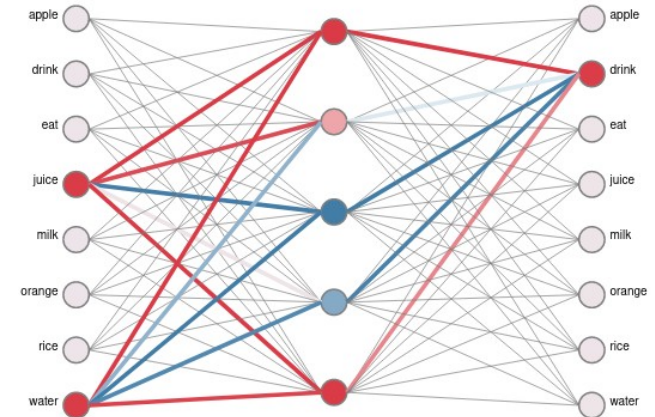
Az MI legújabb reneszánsza: ChatGPT és társai


Gépi tanulással (machine learning) működő








nagy nyelvi modellek (large language models, LLM)

- rendkívül nagy adatkészlet, hatalmas modell, nagyon költséges
OpenAI GPT-3: 45TB szöveg, 800GB modellméret
1 tanítás ~ 4,6 millió USD
- szóbeágyazás (word embedding)
a környezetfüggő jelentés megragadása nagyon tömören
- figyelem (attention)
mire érdemes figyelni a szó környezetéből
- prompt-tervezés: hogyan adjuk meg a feladatokat
„Jelöld meg az alábbi szövegben az eladót és a vevőt!”
„Írj egy programot, ami elkészíti egy szám prímtényező felbontását!”






200,000 United States Dollar equals

70,570,000.00

Hungarian Forint

Jobs based on your profile


58 results




AI Architect, Trilogy (Remote) - \$200,000/year USD

Crossover
Budapest, Budapest, Hungary (Remote)

1 school alum works here

1 day ago · Reposted · **14 applicants** ·  Easy Apply

✕




Head of Data Science

M-KOPA
Pest, Hungary (Remote)

1 week ago · **2 applicants**


✕




Senior AI Engineer, Trilogy (Remote) - \$200,000/year USD

Crossover
Budapest, Budapest, Hungary (Remote)

1 school alum works here

Promoted · **1 applicant** ·  Easy Apply


✕



Lead AI Engineer, Trilogy (Remote) - \$200,000/year USD

Crossover
Budapest, Budapest, Hungary (Remote)

1 school alum works here

Promoted · **0 applicants** ·  Easy Apply

✕

AI Architect, Trilogy (Remote) - \$200,000/year USD

Crossover · Budapest, Budapest, Hungary Reposted 1 day ago · **14 applicants**

Remote · Full-time · Mid-Senior level

5,001-10,000 employees · IT Services and IT Consulting

1 school alum works here

See how you compare to 14 applicants. [Try Premium for free](#)

Skills: Software Development, Artificial Intelligence (AI), +8 more

View verifications related to this job post. [Show all](#)

No longer accepting applications

About the job

Crossover is the world's #1 source of full-time remote jobs. Our clients offer top-tier pay for top-tier talent. We're recruiting this role for our client, Trilogy. Have you got what it takes?

At Trilogy, we see generative AI as more than a trend; we consider it the future scaffold of efficient software engineering. We see tools like Copilot, Cursor, and Code Interpreter as game-changers in the software industry, much like the Internet and cloud computing once were. If you're a seasoned software engineer with a keen interest in AI technologies, this role offers you a unique opportunity to blend your skills in both domains.

Kontextuális szóbeágyazások

- **Statikus szóbeágyazás**
 - a szót (token) minden környezetben azonos vektorral reprezentáljuk
 - probléma: több jelentéssel bíró szóalakok (pl. vár, dob)
- **Kontextuális (dinamikus) szóbeágyazás**
 - a szavak vektorreprezentációját kontextusfüggővé tesszük
 - nagyon nagy korpuszon vizsgálunk hosszú n-gram sorozatokat
- **Hogyan állítsuk elő a modellt?**
 - szavak helyett érdekes részekkel (tokenekkel) dolgozni (gyakoribbak)
 - nem egyformán fontos a kontextus minden része („figyelem” – attention)
 - több absztrakciós szint (modellréteg): szintaxis → szemantika
 - alkalmazott gépi tanulás modellek: LSTM és transzformer hálók
 - eszközök: BERT, ELMo, Flair, GPT stb.
- **Nagy nyelvi modellek (large language models)**
 - kontextuális szóbeágyazások tanítása gigantikus (többnyelvű) adathalmazon

NLP és gépi tanulás: eszközök kísérletezéshez

- **Spacy**
 - teljes NLP feldolgozólánc
 - entitásfelismerés, szófaji címkézés, klasszifikáció, függőség-elemzés
 - [előre tanított modellek](#) (főleg angol)
 - továbbtanítás (adott feladatra)
- **Wevi**: szóbeágyazás-vizualizáció
 - statikus szóbeágyazások
 - böngészőben fut, egyszerű demonstrációk
- **Transzformer modellek**
 - kontextuális szóbeágyazás
 - Hugging Face [intro](#) és [gyakorlatok](#)
- **BERT**
 - transzformer architektúra + előre tanított modellek
 - magyar modellek: [huBERT](#), [emBERT](#), [HILLBERT](#) stb.
- **ChatGPT és OpenAI GPT API**
 - regisztráció, ingyenes próba (API is)
 - generatív modell
nem csak generatív feladatokra
 - Hogyan adjuk meg a feladatot?
„*prompt engineering*”
pl. [entitásfelismerés](#), [hangulatelemzés](#), [webes keresés](#)
 - készülnek [magyar modellek](#) is
pl. NYTUD [Puli](#)
 - [korlátai](#), problémái
 - ár (továbbtanítás)
 - hallucinációk (Solt vs. Debrecen)
 - magyarázatgenerálás hiánya
 - tanítóminta torzítások (bias)
 - adat- és jogvédelem, biztonság
- **Érdekességek**
[Forefront](#) és [character.ai](#)
[Hogyan detektáljuk?](#) [ZeroGPT](#)