# Bemutatkozás, bevezetés

A számítógépes nyelvészet alapjai – ELTE, 2022/23 tavasz 1. óra

Simon Eszter 2023. február 27.

## Tartalom i

- 1. Bemutatkozás
- 2. A félév bemutatása
- 3. Bevezetés a számítógépes nyelvészetbe
- 4. Kis történeti áttekintés

Az MI-kutatás kezdetei

5. Módszerek

Szabályalapú és statisztikai metodológia Neurális fordulat

6. Alkalmazási területek

Információkinyerés, webbányászat Természetes ember-gép kommunikáció Nyelvalapú diagnosztika

#### Tartalom ii

Többnyelvűség, a nyelvi korlátok leküzdése

A nyelvi kulturális tartalmak digitalizálása

7. Házi feladat

# Bemutatkozás

# BEMUTATKOZÁS

- ·én
- ti

# A félév bemutatása

#### ALKALMAK

#### Alkalmak:

- febr. 27.
- · márc. 6.
- márc. 13.
- márc. 20.
- márc. 27.
- ápr. 3.
- · ápr. 10. húsvét
- ápr. 17.
- ápr. 24.
- · máj. 1. munka ünnepe
- máj. 8.
- máj. 15.
- · máj. 22.

Összesen: 11 alkalom

# TELJESÍTÉS FELTÉTELE

- · összesen 5-8 házi feladat kerül kiadásra
- ebből legalább 3-at kell beadni a teljesítéshez (többet is be lehet, ilyenkor a 3 legjobb számít)
- a feladatokból legalább 3 megoldható programozási tudás nélkül is
- · a beadott házikra megajánlott jegyet lehet kapni
- · akinek ez nem jó, jöhet vizsgázni

#### **TEMATIKA**

- 1. Bemutatkozás, bevezetés
  - bemutatkozás
  - · a félév bemutatása
  - · adminisztratív és technikai részletek: git repó, google colab
  - hevezetés
  - · történeti áttekintés
  - alkalmazási területek
- 2. Kis technikai bevezető próbálkozás
  - · Google Colab
  - · python
  - · shell
  - regexek
- 3. Automaták, nyelvtanok
  - · Chomsky-féle nyelvhierarchia
  - automaták
  - · morfológiai elemzés transzducerekkel

#### TEMATIKA - FOLYT.

- 4. Bevezetés a korpuszok csodálatos világába
  - · mi a korpusz?
  - korpusztipológia
  - · mire jó a korpusz?
  - · főbb kérdések a korpuszépítésnél
  - · a korpusz mérete
  - · korpuszannotáció
  - · gyakorlat: crawling, scraping, boilerplate removal
- 5. Korpuszannotáció, annotációs szintek
  - · annotációs eszközök
  - · annotációk összevetése
  - · annotációs szintek:
    - · tokenizálás, mondatra bontás
    - · morfológiai elemzés
    - · morfológiai egyértelműsítés
    - szintaktikai elemzés
    - · szekvenciális címkézési feladatok

#### TEMATIKA - FOLYT.

- 6. Szintaktikai elemzés
  - · konstituenselemzés
  - dependenciaelemzés
  - sekély szintaktikai elemzés
  - · magyar nyelvű elemzők és erőforrások
- 7. Gépi tanulás áttekintés
  - · racionalista és empirikus megközelítés
  - · felügyelet nélküli és félig felügyelt tanulás
  - · a felügyelt gépi tanulás menete
  - · gyakorlat: íriszosztályozás, huntag futtatás
- 8. Vektorszemantika és szóbeágyazások
- 9. n-gram nyelvi modellek

#### TEMATIKA – FOLYT.

- 10. Neurális hálók és neurális nyelvmodellek
  - bevezetés
  - · történeti áttekintés
  - · units
  - · the XOR problem
  - · feedforward neural networks
  - · training neural nets
  - · neural language models
- 11. lauf vagy meghívott előadó

# AJÁNLOTT IRODALOM

- Dan Jurafsky, James H. Martin: Speech and Language Processing.
   3rd ed. draft: https://web.stanford.edu/~jurafsky/slp3/
- Lüdeling, A. and Kytö, M., editors (2008). Corpus Linguistics. An International Handbook. Walter de Gruyter, Berlin
- Mitkov, R., editor, The Oxford Handbook of Computational Linguistics. Oxford University Press, New York.

## TECHNIKAI RÉSZLETEK

## Hozzatok gépet!

https://colab.research.google.com

Minden elérhető lesz a kurzus GitHub repójában: https://github.com/esztersimon/nlp\_at\_elte

Levlista: kérem az emailcímeket!

# Bevezetés a számítógépes

nyelvészetbe

#### SZINONIMÁK?

- · számítógépes nyelvészet
- · természetesnyelv-feldolgozás (natural language processing, NLP)
- · nyelvtechnológia (human language technology, HLT)
- korpusznyelvészet

## BESOROLÁS



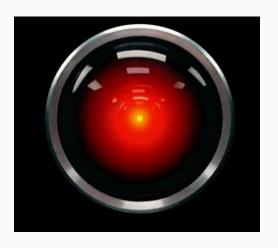
## DEFINÍCIÓ

- · átfedésben van a mesterségesintelligencia-kutatással
- · a természetes nyelvek számítógépes feldolgozásával foglalkozik
- · a kutatások a nyelv szerkezetének gépi modellezésére irányulnak

## Wikipédia:

A számítógépes nyelvészet olyan műszaki tudomány, amely a természetes nyelvű szövegek számítógépes feldolgozásával foglalkozik, de minden olyan elméleti és gyakorlati tevékenység ide tartozik, amely kapcsolatban van a természetes nyelvekkel. Egy interdiszciplína, vagyis olyan szakterület, amely több terület eredményeire és tudására épül, mint pl. az informatika, a matematika és a nyelvészet.

# A SZENT GRÁL



# A NYELVTECHNOLÓGIA CÉLJA

olyan rendszer építése, amely fel tudja dolgozni és elő tudja állítani az emberi nyelvet – úgy, ahogy az ember teszi

elméleti motiváció: az emberi nyelvhasználatot leíró formalizált és konzisztens nyelvi modellek létrehozása

gyakorlati motiváció: a modellek gyakorlati, számítógépes megvalósítása → praktikus gépi alkalmazások

# Praktikus gépi alkalmazások

- · személyi asszisztensek: Siri, Alexa, Cortana, Google Assistant
- · auto-complete
- · spell checking: böngészők, editorok, programok (Microsoft Word)
- · gépi fordítás: Google Translate, DeepL
- · chatbotok, ChatGPT
- · szentimentelemzés (pozitív, negatív és semleges értékelések)
- google calendar bejegyzés emailekből

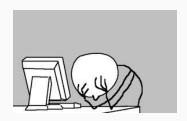
## RÉSZTERÜLETEK

# a nyelvtechnológia egyes részfeladatai tükrözik az emberi nyelvértés pszicholingvisztikai részfeladatait

- · beszédfelismerés -és szintézis
- · morfológiai és szintaktikai elemzés
- · szemantikai elemzés
- generálás
- következtetés

## A PROBLÉMÁK

- · a nyelvfeldolgozás rendkívül bonyolult
- · a szükséges tudás hatalmas
- · szabályalapú: a szabályok száma, a lexikon mérete
- statisztikai: az adatok ritkasága ("rare words are very common")
   → a 15 leggyakoribb szó adja a szöveg 25%-át, a 100
   leggyakoribb a 60%-át, 1000 a 85%-át, 4000 pedig a 97,5%-át
- többértelműség
- magasabb szintű feldolgozási problémák (előfeltevések, mondatok közötti anaforafeloldás stb.)
- robusztusság



# Hogy állunk az egyes részterületeken?

#### nagyon jól:

- · spamszűrés
- · POS-taggelés
- névelemfelismerés (NER)

## egész jól:

- · szentimentelemzés
- koreferenciafeloldás
- jelentésegyértelműsítés (WSD)
- · mondatelemzés, parsing
- · gépi fordítás (MT)
- információkinyerés (IE)

## még mindig nem valami jól:

- kérdésmegválaszolás (QA)
- kivonatolás
- · dialógus

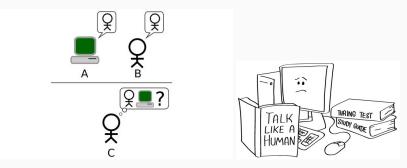
# Kis történeti áttekintés

# Történelemkönyv

- · 1950-60: az első ötletek
- · 1960-70: kísérletezés
- · 1970-80: használható gépek
- · 1980-90: növekvő kapacitás, termékek
- · 1990-: új technológiák, kommunikáció
- · 2000-: növekvő szövegmennyiség, ipar
- · 2010- : neurális fordulat

#### **TURING-TESZT**

- három résztvevő: két tesztalany egy ember és egy gép és egy kérdező
- a kérdező billentyűzet és monitor közvetítésével kérdéseket tesz fel a két tesztalanynak
- mindkét tesztalany megpróbálja meggyőzni a kérdezőt arról, hogy ő gondolkodó ember
- ha a kérdező öt perces faggatás után sem tudja megállapítani, hogy melyik a gép, akkor a gép átment a teszten

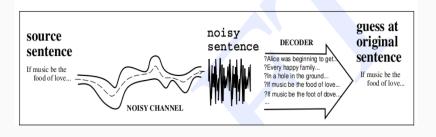


# A TURING-TESZT KRITIKÁJA

- a párbeszéd szimulálása csak kevéssé tekinthető az intelligencia jelének → a hagyományos értelemben vett intelligenciának csak egy szegletét tudja mérni;
- attól még lehet intelligens egy gép, hogy nem képes emberi módon kommunikálni;
- az emberek közül se teljesítené mindenki sikerrel a Turing-tesztet (kisgyerekek, fogyatékosok), holott ők is lehetnek más tekintetben intelligensek;
- a teszten olyan ember is megbukhat, aki nem hajlandó a feltételek szerint együttműködni → az együttműködés megtagadása nem egyenlő az értelem hiányával (lásd HAL);
- a kísérleti szituáció jellegénél fogva a lehetséges beszélgetésfolyamat-variációk száma korlátozott → egy kellően kiterjedt adatbázissal ellátott számítógép előre eltárolt kérdésés válaszminták felhasználásával tényleges intelligencia hiányában is sikerrel teljesítheti a tesztet (lásd Jeopardy)

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3):379–423.

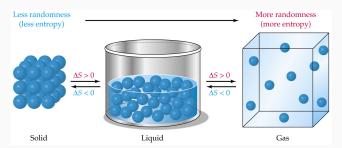
a természetesnyelv-feldolgozási problémák megfeleltethetők dekódolási problémáknak a zajos kommunikációs csatornában



# Shannon, C. E. (1951). Prediction and Entropy of Printed English. Bell Systems Technical Journal, 30:50–64. kikölcsönözte az entrópia fogalmát a termodinamikából, és a

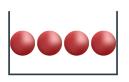
kikölcsönözte az entrópia fogalmát a termodinamikából, és a csatorna információs kapacitásának a mérésére alkalmazta → az információelmélet alapjai

a termodinamikai entrópia egy rendszer rendezetlenségi fokát jellemzi



# AZ INFORMÁCIÓELMÉLETI ENTRÓPIA

- az entrópia akkor a legkisebb (0), ha a hírforrás biztosan mindig ugyanazt a hírt sugározza → a bizonytalanságunk nulla, vagyis teljesen biztosak lehetünk benne, hogy az adott hír fog érkezni
- az entrópia akkor a legnagyobb, ha az összes hír valószínűsége egyenlő → ekkor a bizonytalanságunk a legnagyobb, hiszen bármelyik hír ugyanakkora valószínűséggel érkezhet



High Knowledge Low Entropy



Medium Knowledge

Medium Entropy



Low Knowledge
High Entropy

Kornai videó

# A Georgetown-IBM kísérlet (1954)

- teljesen automatikus gépi fordítás
- · több mint 60 orosz mondatot képes angolra fordítani
- szabályalapú, szótáralapú (a szavakhoz spec. szabályok kapcsolódnak)
  - Operation 0 An exact equivalent for a translated item exists. Any further steps needed.
  - Operation 1 Rearrangement of the position of the words. AB > BA
  - Operation 2 The several choices problem. The result is based on the consecutive words (maximum of three).
  - Operation 3 Also several problems. But the result depends on the previous words (maximum of three).
  - Operation 4 Omissions of the lexical (morphological) item. The source item would be redundant.
  - Operation 5 Insertion of the lexical (morphological) item. The item is not present in the output language.

# A gépi fordítás

- az 50-es évek nagy slágertémája
- · The spirit is willing but the flesh is weak.

$$\rightarrow$$
 orosz  $\rightarrow$  angol  $\rightarrow$ 

The vodka is excellent but the meat is rotten.

- ALPAC (Automatic Language Processing Advisory Committee) report (1966): szkeptikus az eddigi eredményekkel kapcsolatban, az USA kormánya drámaian csökkenti a finanszírozást
- statisztikai módszerek
  - · kapacitásnövekedés  $\rightarrow$  egyre több szöveg válik elérhetővé, kereshetővé
  - sparse data problem: mindig lesznek olyan jelenségek, melyek megfelelő működésének kimutatásához nem elegendő az aktuális méretű adathalmaz
- neurális hálók, vektoros reprezentációk, mélytanulás
  - 2010-es évek: grafikus processzorok (GPU): hatékonyabb számítási teljesítmény + rengeteg adat

#### CHOMSKY

Chomsky, N. (1957). Syntactic Structures. Mouton, The Hague. Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. Language, 35(1):26–58.

Újradefiniálta a nyelvészet feladatát: a nyelvésznek nem a nyelvi jelenségek leírása a feladata, hanem annak a vizsgálata, hogy hogyan tanulja meg a gyerek a nyelvet, és mik azok a jegyek, amelyek minden nyelvben közösek. Márpedig ezek a jelenségek a nyelv felszíni megjelenésétől igen távol esnek, így a "sekély" korpuszalapú módszerekkel nem elérhetőek.

## **STUDENT (1964)**

- · Daniel G. Bobrow
- · középiskolai szöveges matekfeladatokat olvas és old meg
- · bemenet: angol szöveg, kimenet: szám
- pl. ilyeneket: If the number of customers Tom gets is twice the square of 20% of the number of advertisements he runs, and the number of advertisements is 45, then what is the number of customers Tom gets?
- · a mesterséges intelligencia egy korai példája
- · szabályalapú, logikai következtetés

# **ELIZA (1966)**

- · Joseph Weizenbaum, MIT Artificial Intelligence Laboratory
- · az első chatbot
- mintaillesztés, behelyettesítés

```
Welcome to
                                                  AAAAA
                                                   ΔΔΔΔΔΔΔ
                                        ZZ AA AA
777777 AA AA
  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.
ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
       Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
```

### **ELIZA**

- A férfiak mind egyformák.
- · Ezen mit értesz?
- Mindig cseszegetnek 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.

### ...ÉS AMI MÖGÖTTE VAN

### rogersiánus pszichológia

### Carl Rogers (1902-1987):

- · amerikai pszichológus
- · a pszichoterápiás kutatás egyik alapító atyjának tartják
- · kliensközpontú terápia:
  - · a terapeuta párbeszédbe lép a klienssel
  - · bólint, összegzi a hallottakat, ha a másik elakad
  - · a feltárás után továbblép
  - · nem kérdez, figyel

#### Példák

- · egyszerű kulcsszavak által aktivált utasítások: *my boyfriend* 
  - ightarrow your boyfriend
- reguláris kifejezések: s/.\* (depressziós|szomorú)
   vagyok .\*/Sajnálom, hogy \1 vagy/

### AZ ELSŐ KORPUSZOK & A SZTOCHASZTIKUS PARADIGMA

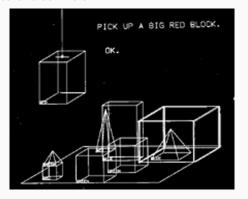
- Brown Corpus (Kucera and Francis, 1967): was created in the US, which then inspired a whole family of corpora:
  - Lancester-Oslo-Bergen Corpus (Leech et al., 1983) (Brown's British English counterpart)
  - · London-Lund Corpus (Svartvik, 1990)

#### A sztochasztikus módszerek

a beszédfelismerés területén érték el az első sikereket, aztán onnan terjedtek tovább más NLP területekre, pl. POS taggelés (Bahl and Mercer, 1976).

### **SHRDLU (1970)**

- Terry Winograd, MIT
- nyelvfeldolgozó, interakció a userrel angol kifejezéseken keresztül
- · memória, statika, névadás
- · az első interakciós fikció



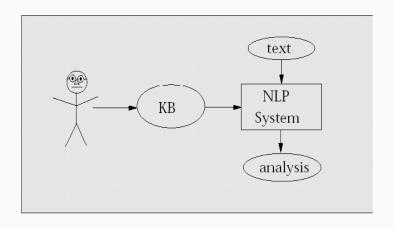
## Chatbotok, asszisztensek és az 1 millió dolláros főnyeremény



- · 2006: Watson (IBM): 2011-ben megnyeri a Jeopardy!-t
- 2011: Siri (Apple)
- · 2014: Cortana (Microsoft), Alexa (Amazon)
- 2016: Google Assistant
- · 2022: ChatGPT

# Módszerek

### SZABÁLYALAPÚ METODOLÓGIA



## Szabályalapú

- © a fejlesztőnek nagy kontrollja van a rendszer fölött
- könnyen értelmezhető visszacsatolás
- magas pontosság
- nyelvi adatok, amik könnyen megragadhatók szabályokkal (reguláris kifejezésekkel), pl. dátumok szerkezete
- © sok kézimunka, nagy szakértelem kell hozzá
- nem hibatűrő
- bonyoult a fejlesztése, törékeny
- nehezen átvihető más doménre, nyelvre
- lehetetlen olyan szabályrendszert írni, ami mindent lefed, amit kell, de semmit, amit nem
- a fedés a listák és a szabályok számának növelésével javítható, de a szabályok száma, a lexikon mérete korlátozott
  - · pl. morfológiai elemzés, tokenizálás

### SZABÁLYALAPÚ METODOLÓGIA

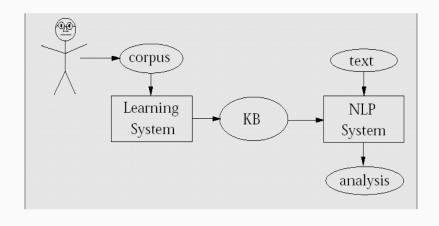
- racionalista filozófiai tradíció (Leibniz, Descartes)
- · univerzális nyelvtan
- velünk született nyelvi képesség → introspekció
- · grammatikalitási ítélet: 0 vagy 1
- kézzel kódolt szabályok
  - reguláris kifejezések

### Példák

e-mail cím:  $[a-z]+@[a-z]+\.[a-z]+$ 

pl.: bubo@doktor.hu

### STATISZTIKAI METODOLÓGIA



## Statisztikai (sztochasztikus), klasszikus gépi tanulás

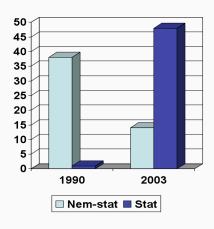
- · adatorientált, gyakorisági adatokból indul ki
- · a nyelv általánosabb megértése, modellálása
- kézzel kinyert feature-ökre támaszkodik (pl. mondathossz, POS-tagek, spec. szavak előfordulása)
- · gépi tanuló algoritmusok (pl. Naive Bayes, SWM, döntési fa stb.)
- · nehézség: az adatok ritkasága ("rare words are very common")
- · pl. szekvenciális címkézési feladatok, szintaktikai elemzés



### STATISZTIKAI METODOLÓGIA

- · empirista filozófiai tradíció (Locke)
- · az érzékszervi tapasztalat prioritása o tudásunk elsődleges forrása a tapasztalat
- · gyakorisági adatokból indul ki, adatorientált
- a szövegből gépi tanuló algoritmus tanulja ki a szabályszerűségeket
- · a grammatikalitási ítélet nem kétértékű, hanem fokozatai vannak

## ÖSSZEHASONLÍTÁS



## ÖSSZEHASONLÍTÁS:)

### Noam Chomsky 1969

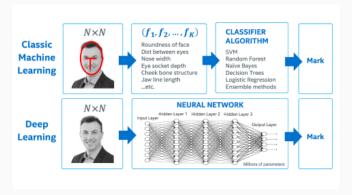
"Meg kell értsük, hogy egy mondat valószínűségéről beszélni teljesen értelmetlen."

#### Fred Jelinek 1988

"Ahányszor távozik egy nyelvész a csoportból, felszökik a beszédfelismerési rátánk."

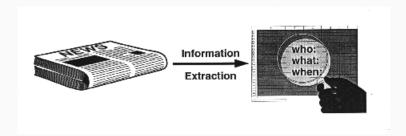
### Neurális

- · 2010-es évek óta ez a legforróbb terület
- nincsenek kézzel kinyert jegyek (self-supervised learning)
- · end-to-end modellek
- · GPU-k, párhuzamosítás, nagyobb számítási kapacitás
- deep learning: azért "mély", mert a neurális hálónak ált. több rétege van



## Alkalmazási területek

## Információkinyerés



az információözönből megtalálni a releváns információt, és csak azt az eredmény strukturált tárolása és megjelenítése

### Részfeladatok

- tokenizálás
- · mondatra bontás
- morfológiai elemzés
- morfológiai egyértelműsítés
- · sekély szintaktikai elemzés
- · tulajdonnév-felismerés
- · koreferenciafeloldás
- · mondaton belüli és mondatok közötti összefüggések felismerése
- szemantikai relációk detektálása

## Felhasználási lehetőségek

- álláspiaci információkinyerés, trendelemzés
  - álláshirdetések és önéletrajzok automatikus begyűjtése, adatbázisba rendezése  $\to$  gyorsabb és hatékonyabb egymásratalálás
- · klinikai információkinyerés
  - kórházi vizsgálati dokumentumok, zárójelentések feldolgozása 

     hasznos információk, statisztika az egészségügy, a gyógyszeripar
     számára
- hangzó anyagokból történő információkinyerés
  - beszédfelismerés → szöveges átirat → szöveges információkinyerő technikák
- véleménykinyerés közösségi tartalmakból
  - blogok, fórumok, bejegyzések feldolgozása → termékekkel, pártokkal, közszereplőkkel kapcsolatos vélemények detektálása

## Természetes ember-gép kommunikáció



automatikus beszédfelismerés, gépi beszédszintézis

## Nehézségek

- a beszédfelismerésben úgy érhető el kellő mértékű pontosság, ha
  - 1. a beszélő által használt szavakat korlátozzuk, vagy
  - 2. a beszélők számát korlátozzuk
- zajrezisztencia: a háttérzaj növekedésével a szófelismerési pontosság rohamosan csökken
- a magyar beszédtechnológia követő helyzetben van: az elsősorban angolra kidolgozott módszereket követi, de a magyar speciális volta miatt ez nem mindig célravezető

## Felhasználási lehetőségek

- · SMS és e-mail felolvasó szoftverek (látássérülteknek)
- · autós és mobiltelefonos GPS rendszer
- orvosi leletező
- · telefonos hívások kezelése, telefonközpont-irányítás
- · hangalapú személyi asszisztens (pl. Siri)
- kötött tematikájú szöveg felolvasása (pl. meteorológiai jelentés, buszmenetrend, betegtájékoztató)

## Nyelvalapú diagnosztika



## Felhasználási lehetőségek

a kóros eseteket tükröző ún. nyelvi markerek keresése beszédben vagy szövegben

- az Alzheimer-kór korai diagnosztizálása beszédtechnológiai fejlesztésekkel
  - a spontán beszéd egyes paraméterei (szünetek, agrammatikus kifejezések) a rövidtávú munkamemória teljesítményéről árulkodnak
- · beszédképző szervek zavarainak diagnosztizálása
  - gégészeti elváltozások, rák korai stádiumban való felismerése a beszédjelek statisztikai akusztikai feldolgozásával

## Felhasználási lehetőségek

- pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatással
  - a pszichológiai folyamatok a verbális viselkedésben is kódolódnak
     → tartalomelemzés → konfliktuselőrejelzés
- korpuszalapú gyereknyelvi kutatások
  - a tipikus fejlődésű gyerekek nyelvének vizsgálata segít az atipikus nyelvi fejlődésű csoportok nyelvi diagnózisában és a fejlesztés kidolgozásában is

## Többnyelvűség, a nyelvi korlátok leküzdése



## Gépi fordítás

### 1. szabályalapú fordítók

- a forrásnyelvi mondat elemzése (→ köztes nyelvi reprezentáció) → célnyelvi szerkezet
- a ritkán előforduló nyelvi szerkezeteket is képes helyesen lefordítani

#### statisztikai fordítók

 párhuzamos korpuszok alapján keresnek célnyelvi megfelelőket valószínűségi alapon

#### 3. neurális fordítók

szószekvenciák prediktálása nyelvmodell alapján

## Számítógépes lexikográfia

szótárak CD-n o "intelligens szótárak" o a szótárkészítés bizonyos lépéseinek automatizálása nyelvtechnológiai eszközökkel

- korpuszalapú módszertan: referenciakorpusz → lexikográfusi munka
- korpuszvezérelt módszertan: a lexikográfiailag releváns információ szövegekből való kinyerése nyelvtechnológiai eszközökkel
  - párhuzamos korpuszok o szó- vagy frázisszintű illesztés o a fordítási jelöltek rangsorolása valószínűség alapján
  - · összevethető (comparable) korpuszok

## A nyelvi kulturális tartalmak digitalizálása



értékőrzés, értékmentés

## Felhasználási lehetőségek

- · magyar nyelvváltozatok adatbázisa
  - · Magyarországon: a dialektusok eltűnése, határon túl: nyelvvesztés
  - · beszélt nyelvi anyag szöveges átirattal o fonetikai, szociolingvisztikai kutatások
- hang/film/multimédia archívumok szövegtartalom szerinti kereshetővé tétele
- parlamenti beszédek tartalmi kereshetősége, folyó beszédek élő feliratozása
- rokon nyelvek nyelvi erőforrásainak fejlesztése

## Nyelvemlékek digitalizálása



## Nyelvemlékek digitalizálása

egyszerű digitalizálás: a primér adat képként való beszkennelése  $\leftrightarrow$  szöveges adatbázisok: nyelvészeti annotációval ellátva

#### hasznos a kutatóknak:

- széleskörű és kifinomult keresési lehetőségek
- · hatékonyabb adatelérés ightarrow időmegtakarítás
- · következetesség, egységesség
- · nem helyhez kötött
- · többen is dolgozhatnak rajta párhuzamosan

## Nyelvemlékek digitalizálása

### hasznos a nyelvtechnológusoknak:

- az elektronikus formátumok előtti korból származó szövegek 
   nehezebb feldolgozás, mint a mai, eleve elektronikusan születő dokumentumok esetében
- $\cdot$  nem sztenderd nyelvváltozatok o az eddiginél robusztusabb vagy új módszerekre van szükség
- · új módszerek  $\rightarrow$  új kutatási kérdések  $\rightarrow$  együttműködés más tudományterületekkel
- · nyelvtörténészek és nyelvtechnológusok ightarrow történeti korpuszok

### Történeti korpuszok

- Penn Parsed Corpora of Historical English
- Newsbooks at Lancester
- Tycho Brahe Parsed Corpus of Historical Portuguese
- Icelandic Parsed Historical Corpus
- Ómagyar Korpusz
- Történeti Magánéleti Korpusz

## Házi feladat

### Házi feladat

- egy nyelvtechnológiát használó alkalmazás bemutatása kb. 10 percben
- · nem kell diasor vagy handout, csak gépen bemutató
- · néhány háttérinfó (ki fejlesztette, mikor, mire jó stb.)
- · legyen interaktív
- jövő óra elején
- péntekig várom emailben a témát, ha valaki vállalkozik erre: simon.eszterke@gmail.com

"That's all Folks!"