



Prószéky Gábor

A nyelvtechnológia alapjai

2019/2020. tanév, 1. félév

A tantárgy órái 2019-ben

- ☐ 1. óra: szeptember 11.
- ☐ 2. óra: szeptember 18.
- ☐ 3. óra: szeptember 25.
- ☐ 4. óra: október 2.
- ☐ 5. óra: október 9.
- ☐ (Pázmány-nap: október 16.)
- ☐ (A köztársaság napja: október 23.)
- ☐ (Őszi szünet: október 30.)
- ☐ 6. óra: november 6.
- ☐ 7. óra: november 13.
- ☐ 8. óra: november 20.
- ☐ 9. óra: november 27.
- ☐ 10. óra: december 4.
- ☐ 11. óra: december 11.

A tantárgy felépítése

- ❑ Előadás:
szerdánként heti 3 óra, azaz max. 135 perc
menete: két részben, egy (kis) szünettel
kezdet: 13.15, vége: 15.45 (témától függően)
- ❑ Gyakorlatok: heti 2 óra
gyakorlatvezető: Yang Zijian Győző
- ❑ Követelmény:
jelenlét előadáson (80%) és gyakorlaton is
+ aktivitás az órán
+ házi feladatok megoldása
+ 2 zárthelyi dolgozat (pótlása kritikus!)
+ kötelező vizsga az idei (!) előadások anyagából
(a diák az előadások után elérhetek lesznek)

A nyelvtechnológia

az **informatikának** az az ága,
amelynek **nyelvészeti kutatásokon alapuló** eredményei
úgy épülnek be – akár ipari méretekben is –
a különféle számítógépes rendszerekbe,
hogy ezek segítségével a gépek a felhasználók számára
a számítógépes kommunikáció folyamán
a nyelvet jól használó emberéhez hasonló
támogatást tudnak adni.



A nyelvtechnológia „elnevezései”

MT

machine translation

gépi fordítás

CL

computational linguistics

számítógépes nyelvészet

NLP

natural language processing

természetesnyelv-feldolgozás

LE

language engineering

nyelvmérnökség

HLT

human language technologies

humán nyelvtechnológiák

Miről lesz szó ebben a tárgyban?

(A tantárgy vázlatos tematikája)

- ☐ A szövegek kódolása
- ☐ Véges állapotú módszerek a nyelvtechnológiában
- ☐ Szóelemzés és -generálás
- ☐ A szóelemzés szerepe a gyakorlatban
- ☐ A mondatok szerkezete és elemzésük
- ☐ Szemantika, világábrázolás, ontológiák
- ☐ Intelligens szótárak
- ☐ Fordítástámogatás, fordítómemóriák
- ☐ Gépi fordítás
- ☐ Statisztikai módszerek
- ☐ Szóbeágyazás, neurális hálók
- ☐ ...



Források a tantárgy irodalmához

- ☐ **Computational Linguistics**
(<http://www.mitpressjournals.org/loi/coli>)
- ☐ **Natural Language Engineering**
(<https://www.cambridge.org/core/journals/natural-language-engineering#>)
- ☐ **Language and Computers**
(<http://www.brill.com/products/series/language-and-computers>)
- ☐ **Computer Speech and Language**
(<https://www.journals.elsevier.com/computer-speech-and-language>)
- ☐ **Corpus Linguistics and Linguistic Theory**
(<https://www.degruyter.com/view/j/cllt>)
- ☐ **Language Resources and Evaluation**
(<http://www.springer.com/education+%26+language/linguistics/journal/10579>)
- ☐ **Machine Translation**
(<http://www.springer.com/computer/ai/journal/10590>)
- ☐ **Journal for Language Technology and Computational Linguistics**
(<http://www.jlcl.org/?language=en>)



További irodalom

Alapkönyv:

- ☐ Daniel Jurafsky & James Martin. *Speech and Language Processing*.
Prentice-Hall, 2000/2008/2019

Rengeteg minden:

- ☐ ACL Anthology
(<http://aclweb.org/anthology-new/>)

Magyarul:

- ☐ Prószéky Gábor: *Számítógépes nyelvészet*.
Számalk, 1989
- ☐ Prószéky Gábor & Kis Balázs: *Számítógéppel emberi nyelven*.
SZAK, 1999
- ☐ Prószéky Gábor: *A nyelvtechnológia (és) alkalmazásai*.
Aranykönyv, 2005



1.

A nyelvtechnológia története

A nyelvtechnológia „evolúciója”

1950-60: ötletek

(vannak már gépek)

1960-70: kísérletek

(kialakulnak az igények)

1970-80: programok

(megjelennek a „használható” gépek)

1980-90: termékek

(a gépek kapacitása megnő)

1990-től: technológia

(a kommunikációs helyzet megváltozik)

2000-től: ipar

(egyre több a feldolgozandó szöveg)

2010-től: internet

(mindenhol)

A gépi nyelvészet „történelmének” kezdetei

- ☐ Általános tapasztalat: a nyelv változik
- ☐ Ezért: a nyelvészet a 20. századig = történeti nyelvészet
- ☐ A deskriptív nyelvészet „mechanikus segédért kiált”
(ami a „preskriptív” nyelvészethoz is jól jön majd!)
- ☐ A számítógép és a gépi fordítás gondolata egyaránt a világháború hozadéka
- ☐ Booth és Weaver: sifrírozás/desifrírozás (1947-49)
- ☐ Bar-Hillel (1951): „a teljesen automatikus gépi fordítás megvalósítható”
- ☐ A gépi fordítás bemutatkozik (némi kormánytámogatással): Georgetown/IBM
- ☐ Szovjetunió és Kelet-Európa: ott inkább matematikai nyelvészet
- ☐ Magyarországon is megindul a gépi fordítás kutatása

Modern nyelvelméletek és elméletmentes gépi kísérletek

- ☐ Generatív grammatika: Chomsky: *Syntactic Structures* (1957)
- ☐ Probléma: a transzformációk nem invertálhatók
- ☐ A számítógépes elemezni akar elsősorban, és nem generálni
- ☐ Amit inkább használnak: kategoriális (Bar-Hillel 1953),
füzér- (Harris 1955) és függősegi (Tesnière 1957) leírások
- ☐ Bar-Hillel (1959): „a teljesen automatikus gépi fordítás
(FAMT) nem lehetséges”
- ☐ Kis (tudománypolitikai) kitérő: hidegháború és holdraszállási
program (1961)
- ☐ Katz & Fodor (1963): megjelenik a szemantika!
- ☐ Chomsky *Aspects of the Theory of Syntax* (1965): szintaktikai
jegyek (következmény: szabály -> szabályosztály)
- ☐ A gépi fordítási korszak vége (1966): az ALPAC Report (ahol -
igen korán - megjelenik a fordítómemóriák alapgondolata!)

Önálló „gépi nyelvelméletek”?

- ☐ ALPAC-következmény: a számítógépes nyelvészet (computational linguistics: CL) megszületése
- ☐ Woods (1969): Lunar (holdprogram!)
- ☐ Egy korszak-meghatározó melléktermék: Augmented Transition Network (ATN)
- ☐ Winograd (1972): a nyelv procedurális közelítése (SHRDLU)
- ☐ Mesterséges intelligencia → számítógépes pszicholingvisztika
- ☐ A gépi fordítás nagy túlélői:
 - Systran (Toma, 1968 → EC, 1976 & Gachot, 1986)
 - Logos (vietnami háború → Wang/IBM/Sun, 1970)
 - Metal (Texas → Siemens, 1978)
- ☐ Gépi fordítás az USA-n kívül:
 - METEO, Eurotra, DLT és az „5. generációs japán álmom”
- ☐ Új fogalom: természetesnyelv-feldolgozás (natural language processing: NLP)

Új generatív elméletek és új gépi megoldások

- ❑ Chomsky „elejtett fonalának” felszedése (1978): a Bay Area-nyelvtanok - GPSG, LFG, HPSG
- ❑ „Frege számítógépesítése”: logikai szemantika és a „rule-to-rule” hipotézisre épülő gépi fordítás (Rosetta)
- ❑ Winograd nyelvi proceduralitása egy „kvázi-elmélet” formájában: *Language as a Cognitive Process* (1983)
- ❑ Elméleti áttörés: a reguláris nyelvtanok és a véges állapotú átmenethálók „újjászületése”: a kétszintes morfológia (Koskenniemi 1983)
- ❑ Megjelennek az első piaci alkalmazások: helyesírás-ellenőrzés, elválasztás - Macintosh, majd IBM PC (1985)
- ❑ A gépi fordítás „leszáll” a PC-re: PC Logos; Siemens Metal > Langenscheidt T1; Systran + Globalink, Kielikone, ProMT
- ❑ Függőségi és tudás-alapú paradigmák a nyelvtechnológiában: kognitív gépi nyelvelméletek, lexikális szemantika (→ WordNet 1985)

A statisztika „mindenhatósága” felé (és tovább)

- ☐ Chomsky újabb elméleteiből a transzformáció nem tűnik el
- ☐ A statisztika „beszáll” a nyelvi modellezésbe (v.ö. Chomsky)
- ☐ Szövegfeldolgozás a beszédtechnológia alapszereivel(1992)
- ☐ Van elég géppel feldolgozható szöveg: megszületik a korpusznyelvész (1995)
- ☐ Nyelv- és beszédfeldolgozás: „ebből igazi üzletet lehet csinálni!” (2000: L&H, SAIL, majd ScanSoft és Nuance)
- ☐ Kialakul a nyelvtechnológia fogalma és megjelenik a gépi beszédfordítás ígérete (2002)
- ☐ Az IBM mesterségesintelligencia- és nyelvtechnológiai „erő-demonstrációkat” tart: Deep Blue (1997) és Watson (2011)
- ☐ Egy ideig gyakorlati, majd már elvi probléma: még sincs elég adat bizonyos témákhoz és nyelvekhez (sparse data problem)
- ☐ Hibrid megoldások: lehet, hogy az ember is így csinálja?
- ☐ Vagy a szóbeágyazás és a neurális hálók oldják meg a problémát?

... de a korpusz-alapú közelítés sem annyira új, mint gondolnánk

Egy új grammatikai módszer van tehát megjelenőben, mely induktív módon halad, azaz a példákból kiindulva ismeri fel a szabályt. A grammatikát tehát az elolvasott, feldolgozott szövegek alapján építjük, úgy hogy a szabályokat a gép a példák segítségével állítja össze statisztikai következtetések útján. Ezáltal ez a módszer véget vet az előre megadott szabályok mechanikus alkalmazásának, és azt indukcióval pótolja. A szabályok így tárolódnak el a gép memóriájában, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.



... de a korpusz-alapú közelítés sem annyira új, mint gondolnánk

„Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatica. A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.”

Riedl Frigyes: Simonyi kis nyelvtana (1882)