

A magyar nyelv a tudományban
MTA, 2021. március 1.



A magyar nyelv digitális fenntarthatóságáért

Prószéky Gábor
Nyelvtudományi Intézet

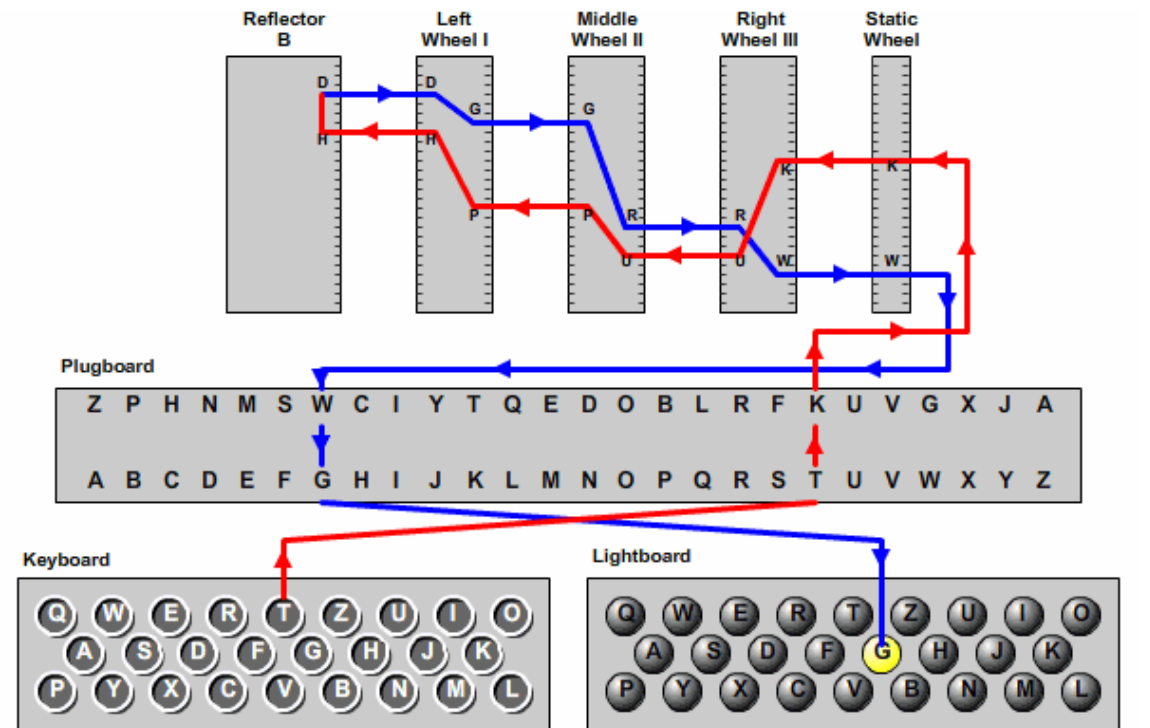
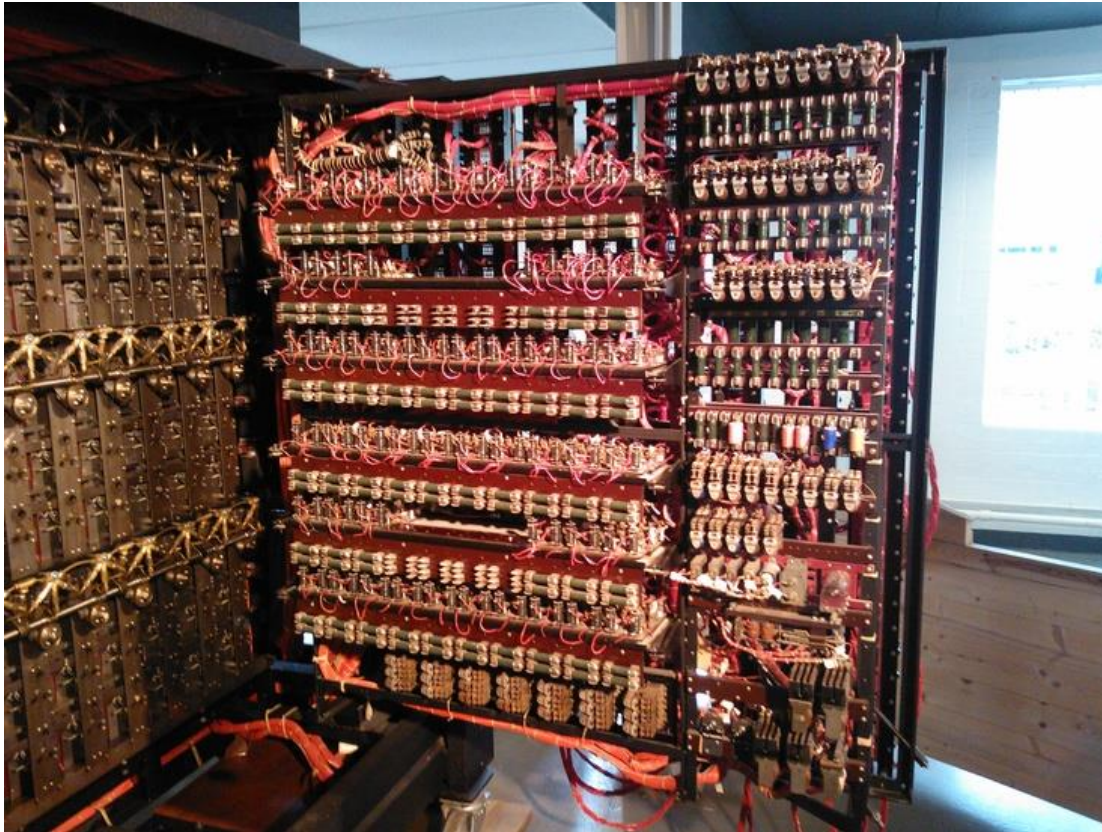
A magyar nyelv digitális fenntarthatósága

- ▶ Manapság a kommunikáció üzleti, társadalmi és egyéni szinten egyaránt digitális eszközökkel és digitális csatornákon, de továbbra is természetes nyelven történik
- ▶ Ezért kulcskérdés, hogy a gépi eszközök egyre nagyobb nyelvi tudással rendelkezzenek, végső célként értsék az üzeneteket (hangban és szövegben egyaránt) és rendelkezzenek természetes nyelvi intelligenciával
- ▶ A nyelvtechnológiai fejlesztés élvonala elsősorban az angolra és néhány nagy nyelvre irányul, ugyanakkor minden fejlett társadalom tagjai joggal várják el, hogy a digitális térben saját anyanyelvüket használhassák, tehát **a különböző nemzeti nyelvek digitális támogatása az egyes országok feladata**
- ▶ Azok a nemzetek, amelyek elmaradnak a nyelvük digitális támogatásával, kiszorulnak a digitális térből, és ez jelentős hátrányt szenvednek egyéni és társadalmi szinten, **ezért a magyar nyelv digitális támogatása nemzetstratégiai kérdés.**

A gépi nyelvészet alaptechnológiái

- ▶ **Szabály-alapú megoldások:** programkódok, nyelvi adatbázisok
- ▶ **Statisztika-alapú megoldások:** kézzel/géppel annotált szövegtörzsek, statisztikai alapszerek, letölthető szoftvereszközök, jelentős háttértárak
- ▶ **Neurális-háló-alapú megoldások:** hatalmas, jó minőségű szövegtörzsek, letölthető szoftvereszközök, kisebb, speciális szakirányú szövegtörzsek, óriási hardverkapacitás (GPU, felhő stb.)

Szabályalapú megoldások (1950-)



© 2006, by Louise Dade

Szabályalapú nyelvi rendszerek

Működő megoldások

Levélszemétszűrés

Let's go to Agra! ✓

Buy VIAGRA ... ✗

Szófaji egyértelműsítés

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Név kifejezés-felismerés

PERSON ORG LOC

Einstein met with UN officials in Princeton

Ígéretes kutatások

Véleménybányászat

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

Koreferencia-feloldás

Carter told Mubarak he shouldn't run again.

Jelentésegértelműsítés

I need new batteries for my *mouse*.

Szintaktikai elemzés

I can see Alcatraz from the window!

Gépi fordítás

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Információkivonatolás

You're invited to our dinner party, Friday May 27 at 8:30

Party May 27
add

Távoli lehetőségek

Kérdés-megválaszolás

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Parafrázis-készítés

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Szövegösszegzés

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialogus

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

A magyar nyelvtechnológia a digitalizálódó Európában

- Basic Language Resource Kit: az ELSNET (European Network of Excellence in Language and Speech) és az ELRA (European Language Resources Association) dolgozta ki
- Magyar Nyelv- és Beszédtechnológiai Platform: elkészül a magyar BLARK
- CLARIN: Common LAnguage Resources and Technology Infrastructure
- Létrejön a HunCLARIN
- CLARIN-ERIC: CLARIN European Research Infrastructure Consortium (25 tag)
- CLARIN az egyik kiválasztott ESFRI Landmark 2016-ban (ESFRI: European Strategy Forum on Research Infrastructures)
- „A magyar nyelv a digitális korban”

A magyar nyelvtechnológia

Kiváló támogatás	Jó támogatás	Közepes támogatás	Töredékes támogatás	Gyenge/semmi támogatás
	angol	csch francia holland lengyel magyar német olasz spanyol svéd	baszk bolgár dán észti finn galíciai görög horvát katalán norvég portugál román szerb szlovák szlovén	ír izlandi lett litván máltai

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE

A MAGYAR NYELV A DIGITÁLIS KORBAN

Simon Eszter
Lendvai Piroska
Németh Géza
Olasz Gábor
Vicsi Klára

■ „A magyar nyelv a digitális korban”

Szabályalapú megoldások a magyar nyelvre

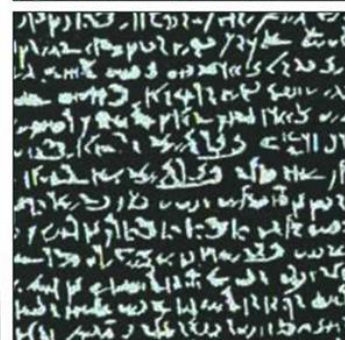
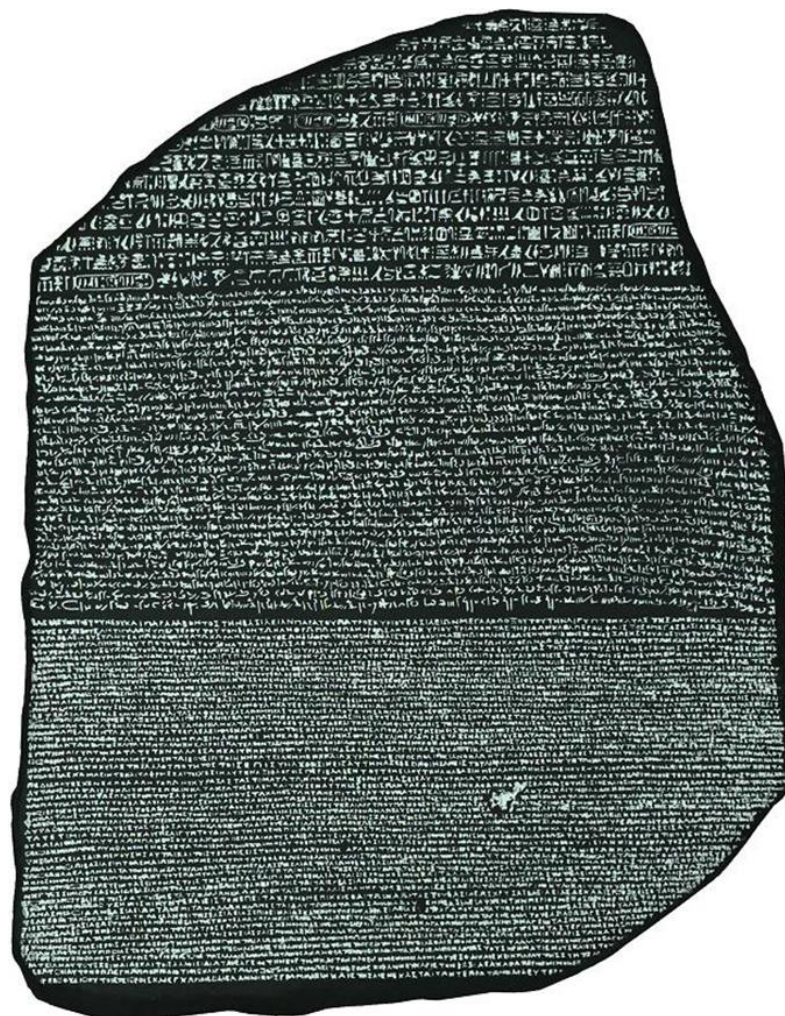
- Morfológiai elemzők: Humor, HunMorph
- Szófaji egyértelműsítők: HunPOS, PurePOS
- Gépi fordító rendszer: MetaMorpho, webforditas.hu

- Termékek írott szövegekhez:
 - Nyelvhelyességi eszközök: Helyesek
 - Intelligens szótárak: MoBiDic, MoBiMouse, MorphoMouse
 - Gépi fordítók: MorphoWord, MoBiCAT

- Az első szövegkorpuszok: Orwell 1984, BUSZI

- Beszédtechnológiai termékek: ProfiVox, MultiVox

Statisztikai nyelvészeti módszerek (1990-)



Statisztika alapú nyelvi rendszerek

Működő megoldások

Levélszemétszűrés

Let's go to Agra!



Buy V1AGRA ...



Szófaji egyértelműsítés

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Név kifejezés-felismerés

PERSON ORG LOC

Einstein met with UN officials in Princeton

Elfogadható megoldások

Véleménybányászat

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



Koreferencia-feloldás

Carter told Mubarak he shouldn't run again.

Jelentésegértelműsítés

I need new batteries for my *mouse*.



Szintaktikai elemzés

I can see Alcatraz from the window!

Gépi fordítás

第13届上海国际电影节开幕...



The 13th Shanghai International Film Festival...

Információkivonatolás

You're invited to our dinner party, Friday May 27 at 8:30



Party
May 27
add

Ígéretes kutatások

Kérdés-megválaszolás

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Parafrázis-készítés

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Szövegösszegzés

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

Dialogus

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?



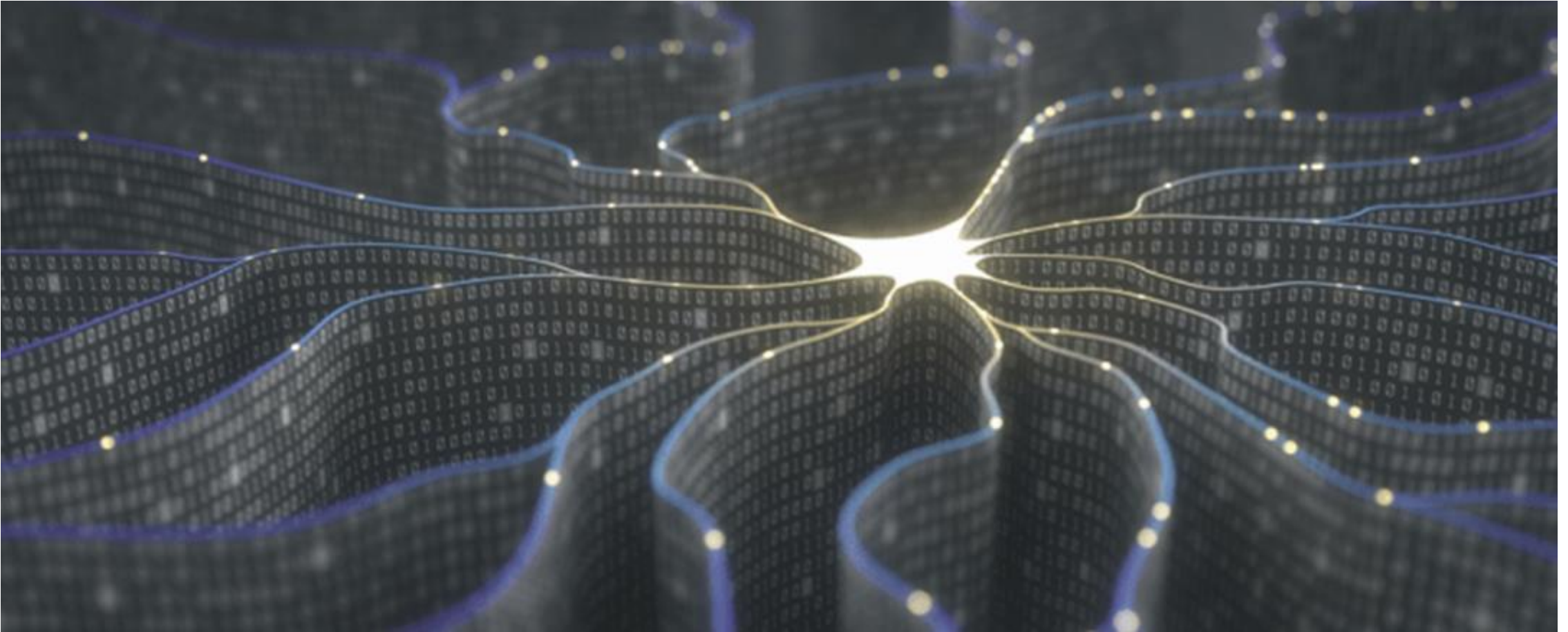
Statisztika alapú megoldások a magyar nyelvre

- Egynyelvű szövegtár: Webkorpusz (589m)
- Annotált szövegtár: Szeged Korpusz (1,2m)
- Kétnyelvű szövegtár: Hunglish (120m/4m mondatpár)
- Pázmány Korpusz (1200m)
- Magyar Nemzeti Szövegtár (187m > 1000m) + Mazsola
- Történeti korpuszok: Ómagyar Korpusz (47 ómagyar kódex, 24 ómagyar szövegemlék, 244 misszilis, 5 középmagyar bibliafordítás: 3,2m), Történeti Magánéleti Korpusz (0,85m), Magyar Történeti Szövegtár (30m)

Statisztika alapú megoldások a magyar nyelvre

- Gépi fordítás – kísérletek (PPKE ITK, EC DGT)
 - Szófaji egyértelműsítés GF-módszerekkel
 - Jelentésegyértelműsítés
 - Névelem-felismerés
-
- Magyar beszédkeltő eszközök, beszédhang-modellek
 - Alkalmazások: közlekedés, telefonos ügyfélszolgálat
 - Beszédkorpuszok (pl. BEA)
-
- Elkészül a – csak részben statisztikai módszereken alapuló – e-magyar.hu

Nyelvfeldolgozás neurális hálókkal (2010-)



Source: <https://www.memsource.com/blog/2017/09/19/neural-machine-translation-the-rising-star/>

Neurális háló alapú nyelvi megoldások

Működő megoldások

Levélszemétszűrés

Let's go to Agra!



Buy VIAGRA ...



Szófaji egyértelműsítés

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Név kifejezés-felismerés

PERSON ORG LOC

Einstein met with UN officials in Princeton

Működő megoldások

Véleménybányászat

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



Koreferencia-feloldás

Carter told Mubarak he shouldn't run again.

Jelentésegértelműsítés

I need new batteries for my *mouse*.



Szintaktikai elemzés

I can see Alcatraz from the window!

Gépi fordítás

第13届上海国际电影节开幕...



The 13th Shanghai International Film Festival...

Információkivonatolás

You're invited to our dinner party, Friday May 27 at 8:30



Elfogadható megoldások

Kérdés-megválaszolás

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Parafrázis-készítés

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Szövegösszegzés

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

Dialogus

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?

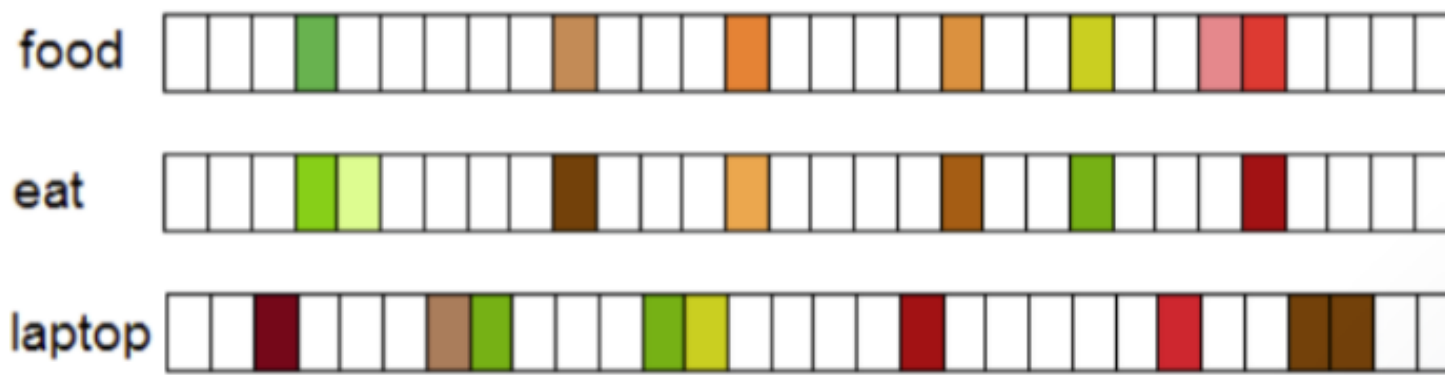


Neurális háló alapú megoldások a magyar nyelvre

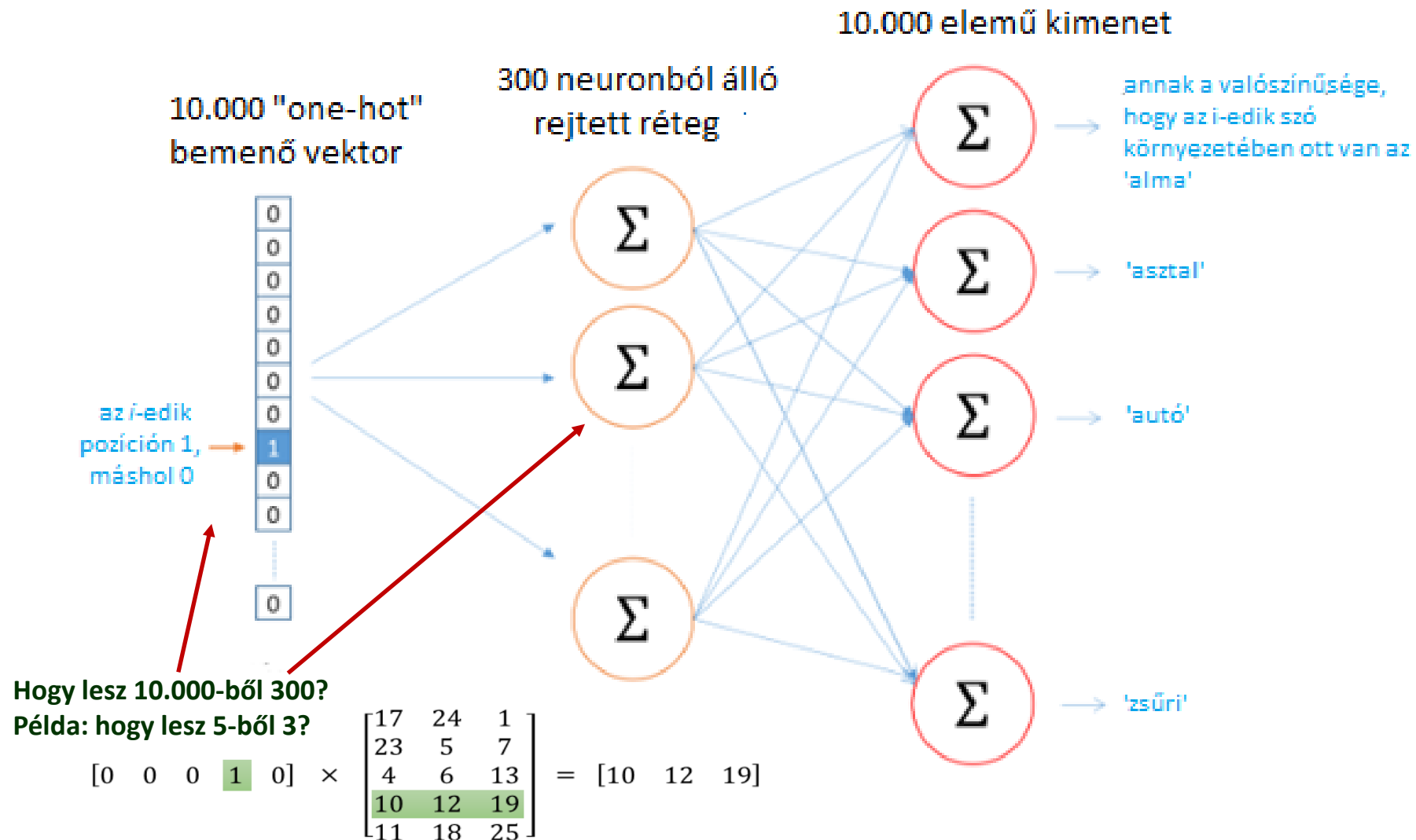
- Szóvektor-reprezentációk: word2vec, fasttext

Szóvektor-reprezentációk

- **Disztribúciós szemantika:** a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusban használjuk őket
- **Szóvektorok:** a lexikai elemek egy valós vektortér egyes pontjai, melyekben az egymáshoz szemantikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, az eltérőek egymástól távol esnek
- A szóbeágyazás adatait nyers **korpuszokból** automatikusan nyerjük
- Két elem szemantikai **hasonlósága** a két vektor távolsága:



Szóvektorok: egyetlen rejtett réteg mint reprezentáció



Jelentéskinyerés szövegtörzsekből

(Novák & Siklósi 2016)

0	kenyerek	1	2270						
1	zsemle	0	Trabantok	1	277				
2	péksütemény	1	Wartburgok	0	csavargó	1	6123		
3	kekszek	2	Skodák	1	koldus	0	rövidnac	1	43
4	pékárak	3	Zsigulik	2	zsivány	1	pizs	0.7731	180
5	tészták	4	Ladák	3	haramia	2	napszemcs	0.7584	37
6	lepények	5	Moszkvicsok	4	csirkefogó	3	sap	0.7460	374
7	kiflik	6	Volgák	5	vadember	4	zacs	0.7259	170
8	kalácsok	7	Daciák	6	útonálló	5	szemcs	0.7209	37
		8	Szukik	7	utcagyerek	6	pih	0.7198	149
				8	gazfickó	7	suzuk	0.6943	131
						8	nemtomm	0.6795	47

Jelentés, forma – és stílus

0	Katalin	1	0	Eufrozina	1	0	Kincső	1	1242
1	Zsuzsanna	0.8893	1	Jolánta	0.7732	1	Csenge	0.8689	4680
2	Ilona	0.8783	2	Konstancia	0.7679	2	Evelin	0.8662	3497
3	Ágnes	0.8750	3	Gertrúd	0.7469	3	Bianka	0.8620	4242
4	Gabriella	0.8735	4	Eugénia	0.7418	4	Fanni	0.8465	10955
5	Judit	0.8730	5	Adelhaid	0.7410	5	Kitti	0.8452	6544
6	Szilvia	0.8483	6	Amália	0.7187	6	Cintia	0.8387	1194
7	Ildikó	0.8465	7	Sarolt	0.7168	7	Villő	0.8358	542
8	Klára	0.8442	8	Gertrud	0.7093	8	Lilla	0.8296	15016

Neurális háló alapú megoldások a magyar nyelvre

- Szóvektor-reprezentációk: word2vec, fasttext
- Dinamikus nyelvmodellek (ugyanahhoz a szóhoz a szöveggörnyezet függvényében különböző vektorokat rendelnek): mBERT, emBERT, HuBERT
- HILBERT: a Nyelvtudományi Intézet és a PTE a Microsoft Hungary támogatásával elkészítette **a magyar nyelv** olyan **digitális nyelvmodelljét** (BERT-large), amelyet **eddig csak tíz nyelvre fejlesztettek ki a világon**
- Finomhangolások/alkalmazások: névelemkeresés, extraktív és absztraktív kivonatolás, szövegosztályozás, ékezetvisszaállítás
- Kutatások: kereső, kérdés-válasz rendszer, fordító
- Termékek: szövegosztályozó rendszer, időpontegyeztető csetbot
- Újabb, nagyobb (címkézett!) szöveggörpuszok, normatív nyelvi anyagok (!)

Neurális háló alapú megoldások a magyar beszédre

- Egyre több beszédkorpusz (BEA, BEKK, MTBA, HuTongue, ...)
- Egyre szélesebb körben használt magyar beszédfelismerők
- Alkalmazások: beszélő gépkönyvek, telefonos ügyfélszolgálati beszélgetések, utólagos leiratozás, valós idejű leiratozás, valós idejű beszédfeliratozás, tematikus diktálás (pl. egészségipar), betegség-előrejelzés
- „Információs akadálymentesítés”

Mi is kellene a biztos folytatáshoz?



Source: <https://www.tmcnet.com/topics/articles/2018/12/26/440734-future-machine-translation.htm>

Amire szükség volna

- A technológia megvan (magyarra is), a hardverháttér megoldható
- Megvalósítás előtt állnak: testre (szakterületre, feladatra) szabott megoldások, amihez abból a „világból” való adatok, szövegtörzsek szükségesek
- Fontos volna a „rejtett internet kinyitása” (szervezeti, jogi korlátok)
- ELRC: célja az ilyen szövegek elérése, technikai és jogi segítség nyújtásával, az EC támogatásával
- A terminológia kulcsszerepben! (de az csak szövegben él)
- Fontos volna a jogi megoldás a különböző szövegek nyelvtechnológiai feldolgozásához való technikai használat biztosítására
- Ennek a jogi környezetéért tudna lobbizni az MTA
- A szakszövegek elérhetővé tételéért az adatgazdák és a központi kormányzati szervek egyaránt tudnak tenni!



Köszönöm a figyelmet!