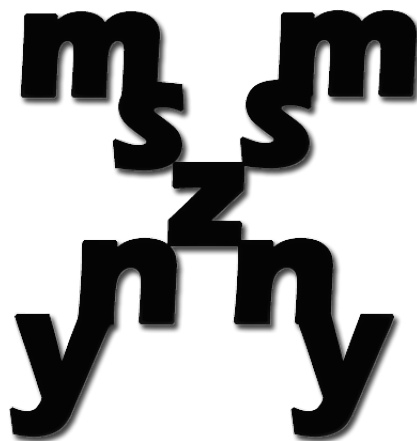


XVII. Magyar Számítógépes  
Nyelvészeti Konferencia



Szerkesztette:  
Berend Gábor  
Gosztolya Gábor  
Vincze Veronika

Szeged, 2021. január 28–29.

**Szerkesztette<sup>1</sup>:**

Berend Gábor, Gosztolya Gábor, Vincze Veronika  
{berendg,ggabor,vinczev}@inf.u-szeged.hu

**Felelős kiadó:**

Szegedi Tudományegyetem  
TTIK, Informatikai Intézet  
6720 Szeged, Árpád tér 2.

**ISBN:** 978-963-306-781-9

**Nyomtatta:**

JATEPress  
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2021. január

**Az MSZNY 2021 konferencia szervezője:**

MTA-SZTE Mesterséges Intelligencia Kutatócsoport

---

<sup>1</sup>a L<sup>A</sup>T<sub>E</sub>X's 'confproc' csomagjára támaszkodva

## Előszó

2021. január 28–29-én már tizenhetedik alkalommal kerül sor a Magyar Számítógépes Nyelvészeti Konferencia megrendezésére. Idén azonban rendhagyó módon, a virtuális térben tartjuk meg konferenciánkat, az ismert COVID-19 járványügyi helyzetre való tekintettel. Ugyanakkor bízunk benne, hogy a személyes találkozások és eszmecsere hiánya ellenére is sikeres és szakmailag mindenkit gazdagító eseménynek nézünk elébe.

A konferencia fő célkitűzése a kezdetek óta állandó: lehetőséget biztosítani a nyelv- és beszédtechnológia területén végzett kutatások eredményeinek ismertetésére és megvitatására, ezen felül a különféle hallgatói projektek, illetve ipari alkalmazások bemutatására. A hagyományokat követve a konferencia idén is nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A 32 beküldött cikkből gondos mérlegelést követően 26-ot fogadott el a programbizottság, melyek témája számos szakterületre terjed ki a beszédtechnológiai fejlesztésektől kezdve a legújabb nyelvi modellek bemutatásán keresztül a spontán beszéd elemzésére vonatkozó eredményekig.

Nagy örömet jelent számunkra, hogy Biszak Sándor és Biszak Előd elfogadták meghívásunkat, akik a digitális archívumok létrehozásával kapcsolatos tapasztalataikról fognak beszámolni plenáris előadásuk során.

Az idei évben is különdíjjal jutalmazzuk a konferencia legjobb cikkét, mely a legjelentősebb eredményekkel járul hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz. Ezen felül immár harmadik alkalommal osztjuk ki a legjobb bíráló díját, amellyel a bírálók fáradtságos, ugyanakkor nélkülözhetetlen munkáját kívánjuk elismerni.

Köszönettel tartozunk az MTA-SZTE Mesterséges Intelligencia Kutatócsoportjának és a Szegedi Tudományegyetem Informatikai Intézetének helyi szervezésben segédkező munkatársainak. Végezetül szeretnénk megköszönni a programbizottság és a szervezőbizottság minden tagjának áldozatos munkáját, ami nélkül nem jöhetett volna létre a konferencia.

A szervezőbizottság nevében,  
Ács Judit, Berend Gábor, Gosztolya Gábor, Novák Attila, Sass Bálint, Simon Eszter, Sztahó Dávid, Vincze Veronika



# Tartalomjegyzék

<b>Nyelvmodellek</b>	<b>1</b>
3	Introducing huBERT <i>Dávid Márk Nemeskey</i>
15	Evaluating Contextualized Language Models for Hungarian <i>Judit Ács, Dániel Lévai, Dávid Márk Nemeskey, András Kornai</i>
29	HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben <i>Feldmann Ádám, Váradi Tamás, Hajdu Róbert, Indig Balázs, Sass Bálint, Makrai Márton, Mittelholcz Iván, Halász Dávid, Yang Zijian Győző</i>
<b>Transzkripció, transliteráció</b>	<b>37</b>
39	Magyar hadifoglyok adatainak orosz-magyar átírása és helyreállítása, és a szabadszöveges adatbázisok tulajdonságai <i>Sass Bálint, Mittelholcz Iván, Halász Dávid, Lipp Veronika, Kalivoda Ágnes</i>
53	emPhon: Morphologically sensitive open-source phonetic transcriber <i>Kulcsár Virág, Lévai Dániel</i>
63	Automatic punctuation restoration with BERT models <i>Nagy Attila, Bial Bence, Ács Judit</i>
<b>Szemantika</b>	<b>75</b>
77	Mitigating the Knowledge Acquisition Bottleneck for Hungarian Word Sense Disambiguation using Multilingual Transformers <i>Gábor Berend</i>
91	Analysing the semantic content of static Hungarian embedding spaces <i>Tamás Ficsor, Gábor Berend</i>
107	Interaktív tematikus-szemantikus térkép a Történeti Magánéleti Korpusz keresőfelületén <i>Novák Attila</i>
<b>Beszédtechnológia</b>	<b>121</b>
123	3D konvolúciós neuronhálón és neurális vokóderen alapuló némabeszéd-interfész <i>Tóth László, Amin Shandiz, Gosztolya Gábor, Zainkó Csaba, Markó Alexandra, Csapó Tamás Gábor</i>

- 139 End-to-end és hibrid mélyneuronháló alapú gépi leiratozás magyar nyelvű telefonos ügyfélszolgálati beszélgetésekre  
*Mihajlik Péter, Balog András, Tarján Balázs, Fegyő Tibor*
- 147 Enyhe kognitív zavar detektálása beszédhangból x-vektor reprezentáció használatával  
*José Vicente Egas-López, Balogh Réka, Imre Nóra, Tóth László, Vincze Veronika, Pákáski Magdolna, Kálmán János, Hoffmann Ildikó, Gosztolya Gábor*
- 157 FORvoice 120+: Statisztikai vizsgálatok és automatikus beszélő verifikációs kísérletek időben eltérő felvételek és különböző beszéd feladatok szerint  
*Sztahó Dávid, Beke András, Szaszák György*

### Spontán beszéd, chat

167

- 169 A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata  
*Vincze Veronika, Üveges István, Szabó Martina Katalin, Takács Károly*
- 183 Magyar nyelvű spontán beszéd szemantikai–pragmatikai sajátságainak elemzése nagy méretű korpusz (StaffTalk) alapján  
*Vincze Veronika, Üveges István, Szabó Martina Katalin*
- 197 Egy nyílt forráskódú magyar időpont-egyeztető chatbot  
*Nagy Soma Bálint, Herdinai Viktor, Farkas Richárd*

### Poszter, laptopos bemutató

209

- 211 StaffTalk: magyar nyelvű spontán beszélgetések korpusza  
*Szabó Martina Katalin, Vincze Veronika, Ring Orsolya, Üveges István, Vit Eszter, Samu Flóra, Gulyás Attila, Galántai Júlia, Svetelszky Zsuzsanna, Bodor-Eranus Eliza Hajnalka, Takács Károly*
- 225 Automatikus írásjelek visszaállítása és Nagybetűsítés statikus korpuszon, transzformer modellel alapuló neurális gépi fordítással  
*Yang Zijian Győző*
- 233 Smooth inverse frequency based text data selection for medical dictation  
*Domonkos Bálint, Péter Mihajlik*
- 243 Automatikus hibajavítás statikus szövegeken  
*Máté Gulás, Yang Zijian Győző, Andrea Dömötör, László János Laki*
- 253 Szó, beszéd – avagy hogyan kommunikálunk egymásról  
*Üveges István, Szabó Martina Katalin, Vincze Veronika*

- 265 Egy következtetésvezérelt csevegőrobot anatómiája. Az ITSy-Bitsy modell  
*Kilián Imre*
- 275 A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei  
*Vincze Veronika, Kicsi András, Főző Eszter, Vidács László*

**Szintaxis, szemantika** **289**

- 291 Jogi szövegek tezaurusz alapú osztályozása: egy nyelvfüggetlen modell létrehozásának problémái  
*Nyéki Bence*
- 305 Egy nagyobb magyar UD korpusz felé  
*Novák Attila, Novák Borbála*
- 319 Értsük meg a magyar entitás-felismerő rendszerek viselkedését!  
*Farkas Richárd, Nemeskey Dávid Márk, Zahorszki Róbert, Vincze Veronika*

**Szerzői index, névmutató** **331**