Bemutatkozik a HunCLARIN

Jelencsik-Mátyus Kinga
MTA Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
Nyelvtechnológiai Kutatócsoport
matyus.kinga@nytud.mta.hu

A HunCLARIN

Számítógépes nyelvtechnológiával foglalkozó tudásközpontok hálózata, akik nemzeti konzorciumot alkotnak.

Jelenleg 8 tagja van, akik

- nyelvi erőforrásokat
- digitális eszközöket
- szakmai segítséget

nyújtanak a nyelvi anyagokkal dolgozó kutatóknak.

Célja

Segítséget nyújtani a bölcsészettudományok, valamint a társadalomtudományok területén nagyobb szövegekkel dolgozó kutatóknak a legfontosabb korpuszok és eszközök (=kutatási infrastruktúrák/KI-k) egyszerű elérhetővé tételével.

A HunCLARIN nemzeti konzorcium tagjai

MTA Nyelvtudományi Intézet - koordinátor

BME Média Oktató- és Kutatóközpont

BME Távközlési és Médiainformatika Tanszék

Szegedi Tudományegyetem

Debreceni Egyetem

MorphoLogic Kft.

Pázmány Péter Katolikus Egyetem

MTA Számítástechnikai és Automatizálási Kutatóintézet

Jelenlegi működés

A HunCLARIN jelenleg egy weboldalon mutatja be a legfontosabb hazai korpuszokat és nyelvtechnológiai eszközöket, valamint a HunCLARIN és a CLARIN híreit, eseményeit.
clarin.hu

Következő lépések

- tárhely (repository) létrehozása az összes HunCLARIN-ban megjelenő erőforrásnak
- ebben jól kereshető metadata és tartalom
- single sign-on felületen keresztül elérhetővé tenni minden tartalmat
- a HunCLARIN megismertetése a szélesebb közönséggel

A single sign-on elérésnek több előnye is van: egyrészt a felhasználónak csak egy helyen kell regisztrálnia és belépnie, másrészt, mivel csak intézményeken keresztül, az ott használt adatokkal lehetséges a belépés, biztosított a felhasználók szűrése, és az adatok (pl. egy beszélt nyelvi korpusz adatközlőinek) védelme.

Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért

2018. október 19. Szeged

A rendezvény támogatói:

Erőforrások

Erőforrás = a HunCLARIN-ban megtalálható kutatási infrastruktúrák (számítógépes nyelvészeti adatbázisok és nyelvfeldolgozó-eszközök)

Jelenleg mintegy 40 KI:

- általános és speciális szövegkorpusz: MNSZ2, Hunglish
- beszélt nyelvi korpuszok: BEA, MONYEK
- multimodális korpusz: HuComTech
- nyelvi feldolgozó eszközök: emMorph, PurePos
- elemzőláncok: e-magyar, magyarlanc

Miért jó a HunCLARIN?

A kutatóknak:

- számos korpuszhoz és eszközhöz hozzáfér egy helyen
- olvashat más kutatásokról, amelyekben HunCLARIN eszközöket használnak
- értesülhet a CLARIN hálózat híreiről, eseményeiről
- segítséget kaphat az erőforrások használatához

A korpuszokat/erőforrásokat megosztóknak:

- adatok biztonságos, hosszú távú tárolása
- többen látják, megismerik és használják az erőforrást

Licenszek

Minden erőforrás a maga által választott lincensznek megfelelő feltételekkel elérhető.
Néhány példa:

- PUB erőforrások - teljesen publikus erőforrások, amelyeket nem korlátoznak személyiségi vagy szerzői jogok (pl. Hunglish korpusz)
- ACA erőforrások - csak kutatási céllal érhetők el, nem kell hozzá engedély, de *federated login*-en keresztül lehet megtekinteni (jelenleg ilyen a Sketch Engine)
- RES erőforrások - számos korlátozást tartalmaz, amelyeket az erőforrás gazdái határoznak meg (pl. BEA korpusz)

HunCLARIN erőforrásokat használó projektek

- **HungaroMars projekt**

  Az MTA TTK-n működő Környezeti Adaptáció és Űrkutatás Kutatócsoportjában, több nemzetközi kutatócsoporttal együttműködve kifejlesztettek egy többnyelvű korpusznyelvészeti pszichológiai tartalomelemző eljárást. Ezt a módszert alkalmazták magyar kutatók egy Mars expedíció szimulációban, valamint ehhez hasonlóan antarktiszi kutatóállomások áttelelő legénységének pszichológiai állapotát is vizsgálták.

- **Diskurzusjelölők a gyermeknyelvben**

  Kondacs Flóra a Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola doktoranduszhallgatója gyermeknyelvi diskurzusjelölő kutatásában a MONYEK-adatbázist is felhasználta saját gyűjtésű korpusza mellett. A MONYEK-adatbázisban található gyermeknyelvi anyagokban először a *hát* diskurzusjelölő funkcióival és a fordulóban betöltött pozícióival, majd az *izé* diskurzusjelölő szerepköreivel foglalkozott.

- **A nyelv koartikulációs működésének vizsgálata**

  Az MTA-ELTE "Lendület" Lingvális Artikuláció Kutatócsoport munkatársai a nyelv mint beszédszerv koartikulációs működésének vizsgálatához használják az elemzőeszközöket (HunMorph, HunToken). Ezek segítségével állítják össze a felvételek anyagát, az adott nyelvi/fonológiai/fonotaktikai szempont szerint keresve a vizsgálandó célszavakat.