A CLARIN és a HunCLARIN

Jelencsik-Mátyus Kinga[1]

[1] Nyelvtudományi Intézet
matyus.kinga@gmail.com

1. Bevezetés

Bár a magyar nyelvet az Európai Unió az erőforrásokkal kevésbé ellátott nyelvek közt tartja számon, a nyelvtechnológiával foglalkozó kutatóközpontokban, egyetemeken, archívumokban mégis jó néhány nyelvi korpusz megtalálható a kisebb speciális korpuszoktól (mint például a BioScope korpusz)[1], a több millió szóból álló írott egynyelvű korpuszokon át (lásd például a Magyar nemzeti szövegtár 2. változatát)[2], az írott és beszélt többnyelvű vegyes korpuszokig (ilyen például az Uráli adatbázis)[3]. Az adatgyűjtemények mellett több magyar kutatóközpontban és egyetemen foglalkoznak nyelvtechnológiai elemzők létrehozásával (lásd például az e-magyar elemzőrendszert)[4]. Ezek a nyelvtechnológiai eszközök jelentősen megkönnyítik nagyobb mennyiségű nyelvi adat feldolgozását a (főként) bölcsészet- és társadalomtudományok kutatásaiban. De még hiányzik egy láncszem: Honnan fognak tudomást szerezni a nem nyelvész kutatók ezekről a lehetőségekről? Kitől kapnak szakmai segítséget a nagy nyelvi adatbázisok feldolgozásához? Sőt, akár a nyelvész kutatók is hogyan fogják megtudni, hogy esetleg más nyelveken vannake már bevált módszerek egy felmerülő probléma megoldására?

2006-ban több mint 20 európai ország nyelvtechnológiával foglalkozó szakemberének részvételével, Váradi Tamás meghívására az MTA Nyelvtudományi Intézetben tartották a CLARIN előkészítő találkozóját. Ez a szervezet épp a fent bemutatott hiányzó láncszem létrehozását tűzte ki céljául.

[1] https://rgai.inf.u-szeged.hu/node/105
[2] http://mnsz.nytud.hu/
[3] http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html
[4] https://e-magyar.hu/hu/

DOI: https://doi.org/10.18135/VT70.13

2. A CLARIN

2.1. A CLARIN célja

A CLARIN (Common Language Resources and Technology Infrastructure) egy európai kutatásiinfrastruktúra-hálózat, amely a digitális nyelvi adatbázisokat és nyelvi feldolgozóeszközöket elérhetővé teszi a bölcsészettudományok és a társadalomtudományok kutatói számára. Kiindulópontja az az elképzelés, hogy az európai és azon túli nyelvek digitális nyelvi erőforrásait egyetlen internetes portálon összefogva egyszerűen hozzáférhetővé tegye. A CLARIN lényegében nem más, mint egy diffúz infrastruktúra, tagintézményekkel (egyetemek, kutatóintézetek) szerte Európában, amelyek szigorú elvárások alapján elnyerhetik a Centre B (K, C, stb.) státuszt.

2.2. Előkészítő szakasz

Két hónappal az MTA Nyelvtudományi Intézetben tartott előkészítő találkozó után benyújtották a CLARIN előterjesztését az Európai Bizottsághoz, majd 2008-ban elkezdődhetett az előkészítő szakasz 22 ország közreműködésével.

Az előkészítő szakasz 36 hónapja alatt megteremtették a megosztott infrastruktúra alapjait. Elsőként kidolgozták az infrastruktúra létrehozásának és működtetésének pénzügyi és irányítási alapelveit, amelyet később az összes részt vevő ország aláírt. A második, kihívást jelentő feladat az addig példa nélküli technikai háttér kialakítása volt, amely lehetővé teszi az összes felmerülő nyelv adatbázisaihoz és nyelvfeldolgozó eszközeihez való egyszerű, egy elérési ponton keresztüli hozzáférést. Harmadikként az infrastruktúra tényleges kialakításához és működésének teszteléséhez a prototípust fel kellett tölteni nyelvi erőforrásokkal minden részt vevő nyelvből. Ebben egyrészt felhasználták a már meglévő korpuszokat és eszközöket, másrészt rávilágítottak arra, hogy számos nyelvben alapvető nyelvi erőforrások is hiányoznak. Ezek létrehozása már a következő szakasz egyik célja lesz. Az előkészítő szakasz negyedik, legfontosabb feladata a felhasználók feltérképezése. Megvizsgálták, mely nyelvtechnológiai folyamatokat használják a leginkább a bölcsész- és társadalomtudományokban. Több kutatásban letesztelték az infrastruktúra használhatóságát. Kiemelten fontosnak tartották, hogy együttműködéseket alakítsanak ki bölcsészek és nyelvtechnológusok között (Váradi és mtsai., 2008).

A szakasz zárótalálkozóját szintén az Intézetben tartották 2011 júniusában.

2.3. Építő szakasz

A CLARIN ERIC (European Research Infrastructure Consortium) 2012-ben jött létre az Európai Bizottság döntése alapján, azzal a céllal, hogy létrehozza és fenntartsa az infrastruktúrát, amely támogatja a nyelvi adatok és eszközök megosztását, használatát és fenntarthatóságát főként a bölcsészet- és társadalomtudományok számára. A CLARIN ERIC-nek tagja lehet ország vagy kormányközi szervezet. Magyarország, bár a kezdetektől jelen volt a folyamatokban, csak 2016. augusztus 1-jén csatlakozott hivatalosan is a konzorciumhoz. A CLARIN-nak jelenleg 21 tagja és 3 megfigyelő státuszú országa van. Az egyes országokon belül a tagok (jellemzően kutatóintézetek, egyetemek, könyvtárak, archívumok) létrehoznak egy nemzeti konzorciumot. A CLARIN tehát egy szétszórt infrastruktúra szerte Európában, ahol a tagok nyelvi korpuszokat, digitális nyelvfeldolgozó eszközöket, valamint szakmai segítséget nyújtanak a nyelvi anyagokkal dolgozó kutatóknak.

Az infrastruktúra gerincét a központok alkotják. Központ lehet minden olyan intézmény vagy nemzeti konzorcium, amely megfelel a szigorú elvárásoknak, és végigmegy az engedélyeztetés folyamatán. A legfontosabb központtípus a B, a szolgáltatást nyújtó központ. Ezek alkotják a CLARIN magját. Ezek a központok olyan szolgáltatásokat nyújtanak, amelyek többek közt hozzáférést biztosítanak az általuk tárolt nyelvi korpuszokhoz, és az általuk kifejlesztett eszközök folyamatosan elérhetőek valamely CLARIN-nak megfelelő felületen.

A K központok tudásközpontok, amelyek szakmai segítséget nyújtanak a kutatónak ahhoz, hogy használni tudják a CLARIN nyújtotta szolgáltatásokat. Az egyes K központok eltérő területeken segítik a kutatókat. A C központok metaadatokat szolgáltatnak folyamatosan elérhető módon. Az E központok külső központok, amelyek a CLARIN-hoz kapcsolódó szolgáltatásokat nyújtanak, de nem a CLARIN tagjai. A CLARIN jelenlegi központjai láthatóak az 1. képen.

1. kép. A CLARIN központjai.[5]

2.4. Üzemeltetési szakasz

Ma körülbelül 20 B, és számos más típusú központ van a CLARIN-ban, számuk folyamatosan növekszik, a szervezet tehát a különböző központok hálózataként működik. A gondos előkészítés után a több éves működés alapján látható, hogy a CLARIN egyszerű és fenntartható hozzáférést nyújt a digitális nyelvi adatokhoz (írott, beszélt vagy multimodális) a bölcsészet- és társadalomtudományok kutatóinak. Fejlett eszközöket biztosít a nyelvi adatok kutatására, elemzésére. Lehetőséget nyújt a nyelvi korpuszok és eszközök kombinálására, összehasonlítására, valamint szakmai segítséget kínál mindezek használatához (Jong és mtsai., 2018). Technikai háttér tekintetében nyelviadat-repozitóriumok, szolgáltató központok és tudásközpontok állnak a részt vevő országok kutatói szolgálatában, egy egyszerű single sign-on eléréssel. Elmondható tehát, hogy az adatok és eszközök interoperabilitása megvalósult (Hinrichs és Krauwer, 2014).

A CLARIN ma számos országban tökéletesen működik. A meglévő korpuszok és eszközök fejlesztéséhez segítséget nyújtanak, az újonnan jelentkező országokban pedig segítik a rendszer kiépítését.

[5] A kép forrása: https://www.clarin.eu/content/overview-clarin-centres

3. A HunCLARIN

A HunCLARIN a vezető hazai nyelv- és beszédtechnológiai kutatásfejlesztést végző tudásközpontok stratégiai jelentőségű kutatásiinfrastruktúra-hálózata (SKI).

A kutatások bázisát képező nyelvi erőforrásokat és eszközöket tartalmaz. A megosztott virtuális hálózat 2010-ben, majd 2015-ben ismét SKI minősítést kapott. A HunCLARIN-hoz eddig 8 partner csatlakozott:[6] Nyelvtudományi Intézet (mint a HunCLARIN központja), BME Média Oktató- és Kutatóközpont, BME Távközlési és Médiainformatikai Tanszék, Szegedi Tudományegyetem, Debreceni Egyetem, Pázmány Péter Katolikus Egyetem, Morphologic Kft., valamint a Számítástechnikai és Automatizálási Kutatóintézet.

Az ezekben a központokban létrehozott jelenleg több mint 40 tag számos általános és speciális szövegkorpuszt, különféle nyelvi feldolgozó eszközöket, elemzőket, adatbázisokat, ontológiákat ölel fel.[7,8] A hálózat koordinátora és kapcsolattartója: Váradi Tamás.

A HunCLARIN legfontosabb célja a tudományos kutatás támogatása a nyelvtechnológia, a nyelvi erőforrások könnyű elérhetővé tételével. Ennek alapfeltétele egy olyan internetes felület, valamint az annak hátterében álló technikai infrastruktúra létrehozása, amelyen keresztül (a regisztrált kutatók számára) a csoportban található összes KI egyszerűen elérhető, valamint az eszközök egymással és a CLARIN más nyelveken megvalósuló alkalmazásaival összevethető. Ezzel lényegesen egyszerűbbé válik a magyar nyelv- és beszédtechnológia bekapcsolása a magas szinten folyó európai munkálatokba, hiszen a CLARIN számos más európai tagjánál (és azok között) a nyelvtechnológiai eszközök és erőforrások interoperabilitása már megvalósult.

A HunCLARIN tagjai számos jelentős hazai és nemzetközi projektben vettek részt. Ilyen például az uráli–orosz kontaktushatás kutatását is lehetővé tevő többnyelvű Uráli adatbázis, amely írott és beszélt nyelvi szövegeket is tartalmaz udmurt, tundrai nyenyec, szinjai és szurguti hanti nyelven.

Ahogy az 1. képen látszik, Magyarországra még nincs központ jelölve, de a HunCLARIN célja a B központ státusz elérése.

---

[6] http://clarin.hu/content/hunclarin-tagjai
[7] http://clarin.hu/content/korpuszok
[8] http://clarin.hu/content/nyelvtechnol%C3%B3giai-eszk%C3%B6z%C3%B6k

4. A felhasználók bevonása

A CLARIN, és vele összhangban a HunCLARIN is nagy hangsúlyt fektet a felhasználók, illetve a leendő felhasználók bevonására, tájékoztatására. Konzorciumon belül, tehát a magyarországi tagok közt, valamint nemzetközi szinten is évente számos alkalommal rendeznek előadásokat, workshopokat és webináriumokat. Ezek során nagy hangsúlyt fektetnek arra, hogy a résztvevőknek lehetőségük legyen kötetlen módon információkat szerezniük.

A CLARIN, illetve a HunCLARIN bemutatásának, valamint a más kutatóközösségekkel való kapcsolatépítésnek egyik nagyon hatékony módja a roadshow. Ezt bizonyítja az eddig megrendezésre került 3 rendezvény is Szegeden, Debrecenben, illetve Pécsett.

A roadshow lényege, hogy házhoz viszi a nyelvtechnológiát oda, ahol a bölcsész és társadalomtudományi kutatások zajlanak, vagyis az egyetemekre. Ezeknek az eseményeknek a szerkezete mindig úgy épül fel, hogy a nap kezdetén a HunCLARIN központból érkező nyelvtechnológusok röviden ismertetik a HunCLARIN, illetve a CLARIN célkitűzéseit, felépítését, működését, majd bemutatják, milyen korpuszokat és nyelvfeldolgozó eszközöket nyújthatnak a kutatók számára. A második részben a helyi bölcsészettudományi műhelyekben zajló munkákba kaphatunk betekintést, amelyekben nyelvtechnológiai eszközöket is igénybe vettek a nyelvi adatok elemzéséhez. Mindkét részben nagy hangsúlyt fektettek a közönség és az előadók közti párbeszédre.

Bibliográfia

Hinrichs, E., Krauwer, S.: The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1525–1231. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)

Jong, F. de, Maegaard, B., De Smedt, K., Fišer, D., Van Uytvanck, D.: CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In: Calzolari, N. et al. (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018. pp. 3259–3264. European Language Resources Association (2018)

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., Koskenniemi, K.: CLARIN: Common Language Resources and Technology Infrastructure. In: Calzolari, N. et al. (eds) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). pp. 1244–1248. European Language Resources Association (2008)