

Magyar korpusznyelvészet és nyelvtechnológia

Váradi Tamás
MTA Nyelvtudományi Intézet
Korpusznyelvészeti osztály
varadi@nytud.hu

Jelen dolgozat áttekintést kíván adni a magyar nyelvű korpusznyelvészet eredményeiről valamint az MTA Nyelvtudományi Intézet Korpusznyelvészeti osztályán folyó nyelvtechnológiai kutató-fejlesztő tevékenységről. Három területet szeretnék érinteni: a magyar nyelvű korpuszok, a párhuzamos korpuszok és a lexikai adatbázisok fejlesztésével kapcsolatos munkákat.

Magyar nyelvű korpusznyelvészet

A magyar nyelvű korpusznyelvészet az Akadémiai Nagyszótár munkálataiból született meg. A már száz éves, meglehetősen hányatott, múltra visszatekintő munkálatok 1984-ben új lendületet kaptak, amikor is az Akadémia határozatban kötelezte el magát a munkák újraindítása mellett. A szótár készítése módszertanilag is megújult. Papp Ferenc javaslatára az Akadémiai határozat kimondta, hogy a szótárat korszerű technológiával, számítógéppel kell készíteni és anyagát számítógépes korpuszra kell alapozni.

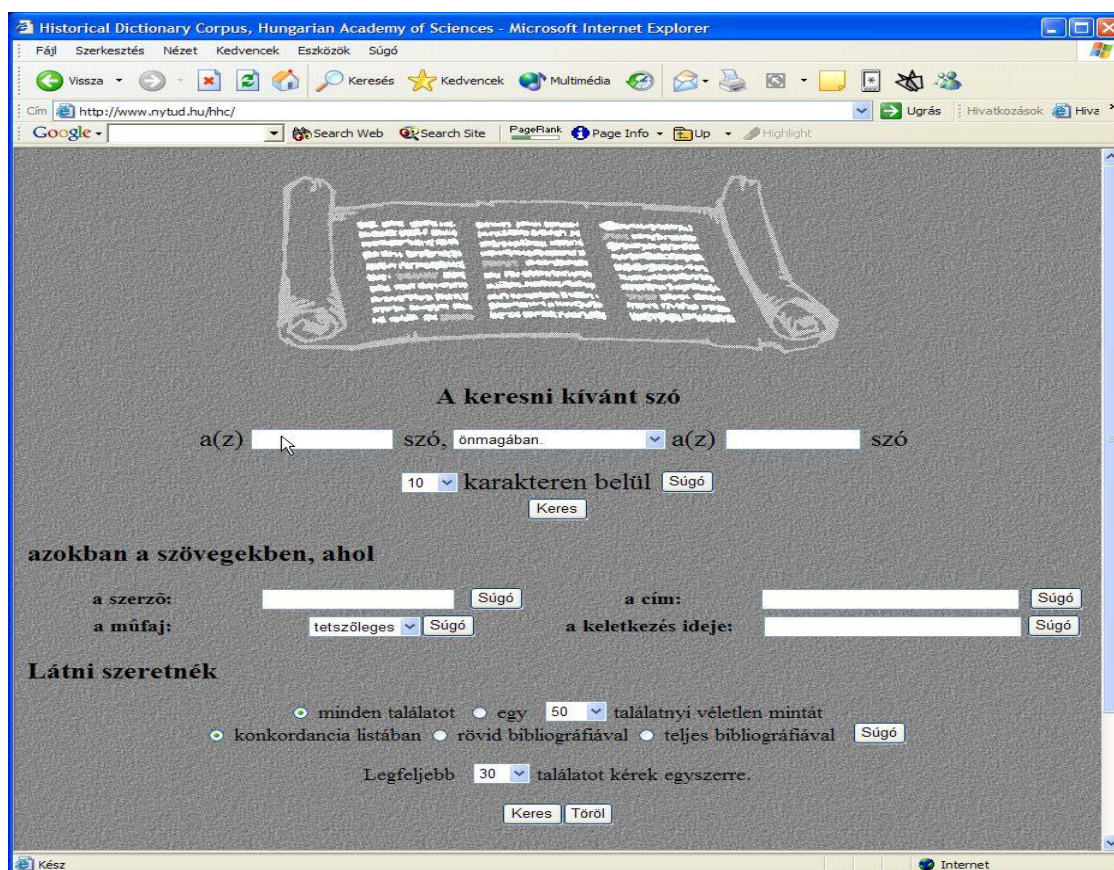
Magyar Történeti Korpusz

Ebből a tervből született meg a Magyar történeti korpusz, amely eredetileg 10 millió szavas összeállítás lett volna 16 – 20. századi szövegekből. A korpuszba bekerülő szövegeket felkért szakértőkből álló bizottságok válogatták ki. A szövegek nagyobb része szépirodalmi alkotás, de az írásos nyelvhasználat egyéb műfajai is tág körben megtalálhatók benne. Hosszabb művekből csak részletek kerültek be, a szövegek hossza még így is széles határok között változik. Minden szöveget Xerox másolatból kézzel, begépelés útján vittek számítógépre.

A nyolcvanas évek második felében megindult korpuszépítési munka a kilencvenes évek közepére lezárult, és az eredeti terveket mennyiségileg jócskán meg is haladta. A korpusz az eredetileg tervezett 10 millió helyett mintegy 23 millió szövegszót tartalmaz, és különféle gyakorlati korlátok miatt a könyvnyomtatás magyarországi kezdete helyett a 18. század elejétől a 20. végéig származó szövegeket ölel fel. Főbb jellemzőiről l. Pajzs–Váradi 1997, Váradi 1997.

A szövegek kódolásakor nemcsak a szöveg bibliográfiai adatait és szerkezeti egységeit (pl. fejezetek, bekezdések, mondatok) rögzítették, hanem minden egyes szövegszó morfológiai elemzését is elvégezték. Ezt a feladatot a Morphologic által kifejlesztett HUMOR morfológiai elemzővel végezték (Prószték – Tihanyi 1996).

A régi korokból származó szövegek feldolgozása különleges nehézségek elé állította a korpusz készítőit. A betűhelyes lejegyzés két technikai problémát is felvetett. A viszonylag könnyebb kérdést az olyan “betűk” pontosabban grafikai alakzatok jelentették, amelyek ma már nem léteznek a magyar ábécében, és természetesen nem találhatók a számítógép billentyűzetén illetve a karakterkészletében sem. Az ilyen alakok kódolására az ún. Prószték-kódkészlet kiterjesztett változatát használják, amely az alapbetű mellé tett szám segítségével kódolja az illető hang variánsait. A nehezebb kérdést az olyan szóalakok jelentetik, amelyek a mai nyelvben nem léteznek: vagy a szó maga kihalt, vagy a toldalék nem használatos ma már. A kihalt alakváltozatok kezelésére külön heurisztikus eljárást kellett kidolgozni (Kiss et al. 2001). A korpusz jelenleg már napi használatban szolgálja az Akadémiai Nagyszótáron dolgozó lexikográfusokat. Nyilvános változata a <http://www.nytud.hu/hhc> weboldalon érhető el. Az 1. ábrán láthatjuk a hálózati kereső felület képét.



A Magyar Nemzeti Szövegtár

A Történeti Korpusz munkálatai befejeztével felmerült az igény, hogy a diakrón korpusz mellett szükség lenne egy nagyméretű, az aktuális nyelvhasználatot tükröző szinkrón korpusz összeállítására is. E nagyméretű vállalkozás hívta létre 1997-ben az MTA Nyelvtudományi Intézetén belül a Korpusznyelvészeti osztályt, melynek központi feladata a ma már Magyar Nemzeti Szövegtár néven ismert korpusz megalkotása lett.

Az eredeti cél 100 millió szó összeállítását irányozta elő, amely a legújabb írásos nyelvhasználatot volt hivatott tükrözni. A korpuszépítés alapkérdése természetesen a szerkezetük meghatározása. Érthetően minden korpusz arra törekszik, hogy reprezentatív mintája legyen az adott nyelvnek, de be kell látni, hogy maga a megcélzott totalitás, azaz a nyelvhasználat egésze, annyira összetett jelenség, hogy ennek szigorú értelemben vett mintáját elkészíteni elvileg is lehetetlen feladat (Váradi 2000). A Magyar Nemzeti Szövegtár öt markánsan elkülönülő nyelvhasználat köré szerveződött: sajtó, szépirodalom, (tudományos) értekező próza, hivatali nyelvhasználat és személyes közlés. Az MNSZ moduláris felépítésű, ami azt jelenti, hogy ez az öt nyelvhasználati kör egy-egy részkorpuszt alkot. A korpusz lekérdezésekor az egyes részkorpuszokat együttesen vagy tetszőlegesen bármilyen kombinációban is bevonhatjuk a vizsgálatba. Az 1. táblázat tartalmazza az MNSZ szerkezetét és méretarányait.

Nyelvi változat	M. szó	Forrás
Sajtó	75	online kiadások
Szépirodalom	15	Digitális Irodalmi Akadémia
Tudományos próza	20	Magyar Elektronikus Könyvtár
Hivatali nyelv	20	minisztérium, önkormányzat
Személyes közlés	20	Index.hu – Törzsasztal
Összesen:	150	

1. táblázat A Magyar Nemzeti Szövegtár összetétele

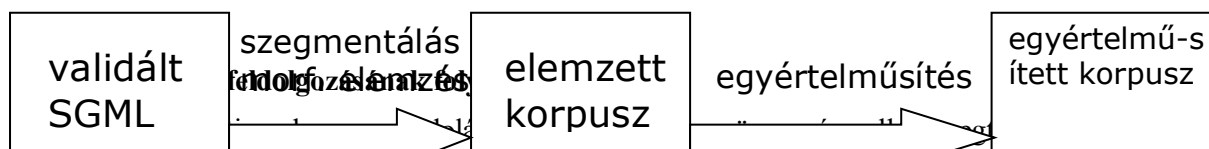
Gyakorlati korlátok miatt az MNSZ-be csak már elektronikus alakban létező szövegek kerülhettek, melyek fő forrása az internet. A munkálatok indulása idején ez nagyon kockázatos döntésnek tűnt, de az idő haladtával olyan rohamos mértékben növekedett az interneten közzétett magyar nyelvű szövegek köre, hogy ma már ez a szempont szinte semmi korlátozást nem jelent a szövegek válogatásában. Ha pedig netán felmerülne az az aggály a forrással kapcsolatban, hogy az internethasználók Magyarországon egyelőre a beszélők körére korlátozódik, akkor emlékeztetnünk kell arra a tényre, hogy a nyomtatásban megjelenő szövegeket valójában egy még ennél is szűkebb réteg állítja elő. Vagyis a helyzet épp fordított: az internet sokkal demokratikusabb publikálási forma, és ezáltal sokkal szélesebb körű mintavételt tesz lehetővé, mint a hagyományos nyomtatott médium.

- A sajtóanyag volt a legkorábban elérhető az interneten, és azóta is ez a forrás ontja a legnagyobb tömegben a szövegeket. Az MNSZ válogatásában szerepelnek az országos napilapok mellett regionális kiadványok, és jelen van a határon túli magyarság sajtójának egy-egy képviselője. Természetesen a címke „sajtónyelv” a szövegváltozatok széles skáláját takarja, hiszen az itt található szövegek között éppúgy van novella, esszé és vezércikk mint tőzsdehírek, receptek és moziműsor.
- A szépirodalmi alkorpusz külön igen értékes részét képezi az MNSZ-nek, mivel ez teljes egészében tartalmazza a Digitális Irodalmi Akadémia anyagát. Ez azt jelenti, hogy az élő magyar klasszikus írók teljes életműve vizsgálható lesz a korpusznyelvészet módszereivel. A közzététel nem jelent egyszerű duplikációt, hiszen, amint látni fogjuk, az MNSZ nyelvileg elemzett korpusz, és ez lényegesen pontosabb lekérdezést tud lehetővé tenni, mint a szöveg felszíni alakjára irányuló keresés.
- Az értekező próza a Magyar Elektronikus Könyvtár társadalom- és természettudományos részéből került ki. A cikkeket, monográfiát tartalmazó válogatás mellett szerepel a Természet Világa folyóirat anyaga is.
- A hivatali nyelv kategória igen vegyes anyagot tartalmaz. Ide tartoznak például különböző jogszabályok, rendeletek, közlemények, melyeket minisztériumok, önkormányzatok, különböző hivatalos szervek tettek közzé. Egy jelentős részét e kategóriának az Országgyűlési Jegyzőkönyvek anyaga teszi ki.
- Az ún. személyes közlés alatt az Index.hu internetes újság Törzsasztal fórumának anyagát foglalja össze a kezdetektől 1999-ig. A Törzsasztal az Index.hu legnépszerűbb, leglátogatottabb fóruma, ahol a témák rendkívül színes kavalkádjá került napirendre, melyek az olvasók egész széles körét késztették hozzászólásra. Ez az alkorpusz azért kapta a személyes közlés elnevezést, mert véleményünk szerint ezek a szerkesztetlen, névtelen hozzászólások közelítik meg a legjobban a spontán vita közben megnyilvánuló természetes nyelvhasználatot. Szociolingvisztikai szempontból is figyelemre méltó e vadonatúj műfaj sajátosságainak vizsgálata.

Az MNSZ nem csupán egységes keretbe foglalja a szöveget, hanem fel is dolgozza őket. Ez a munka egyrészt a szövegek bibliográfiai adatainak jelölését jelenti, amelyet a szövegek elé helyezett bevezetőben találhatunk. A szöveg fő szerkezeti és tartalmi egységeit is jelöljük. Ezen felül minden egyes szövegszó morfológiailag elemzett alakban szerepel. A morfológiai elemzés a MorphoLogic Humor morfológiai elemzőjével készült.

A morfológiai elemzés a szavak közel egyharmadára többféle értelmezést ad. A *falunk* szóalak például egyaránt értelmezhető igeiként valamint a *fal* vagy a *falu* főnév birtokos személyragos alakjaként. Egyedülálló teljesítménye az MNSZ-nek, hogy a korpusz teljes állományában sikerült a szavakat 97,4%-os pontossággal egyértelműsíteni. Ezt az eredményt egy erre a célra kidolgozott statisztikai alapú eljárással sikerült elérni (Oravecz–Dienes 2002). A Szegedi Egyetem által kidolgozott automatikus egyértelműsítő eljárás (1. 2. fejezet) a gépi tanulás módszerét alkalmazta az MNSZ-ével összevethető minőségben (Alexin et al. 2001). Mindez a sok mögöttes információ az SGML jelölőnyelv használatával a nyelvi korpuszokra kidolgozott CES szabvány szerint (Ide 1999) van a szövegbe beültetve. A 2. ábra a korpusz feldolgozásának egyszerűsített folyamatát ábrázolja.





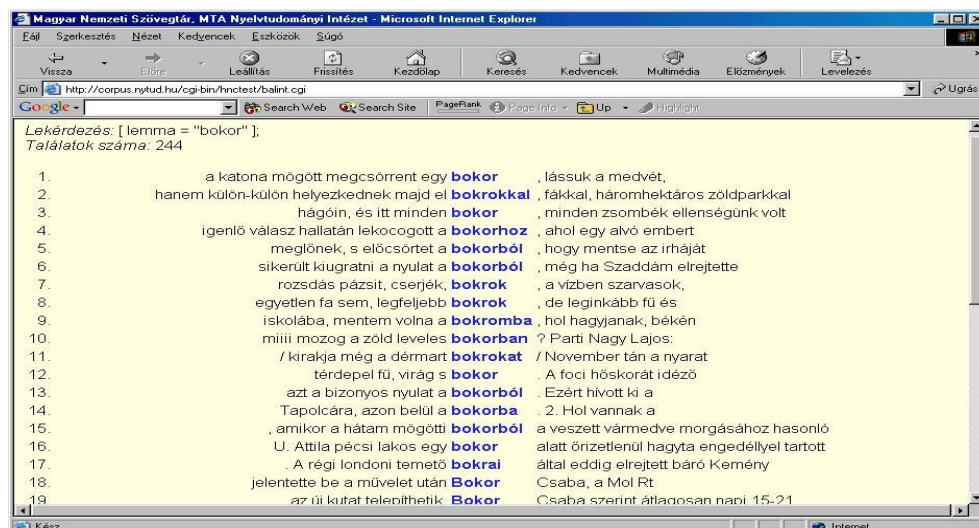
szóalak morfoszintaktikai jellemzői (*msd*) valamint a korpusz további feldolgozására szolgáló címke (*ctag*). A lekérdezésben a szótő, a szóalak, és a nyelvtani jellemző egyaránt megadható, ami lényegesen növeli a hatékonyságot. Ezt két rövid példában láthatjuk.

Ha a szótőt adjuk meg a keresés kiindulópontjának, akkor biztosak lehetünk benne, hogy az illető szó valamennyi toldalékolat alakját megkapjuk még akkor is, ha a toldalékolás egyben a fő változásával is jár. A 4. ábrán a *bokor* szó szótő szerinti lekérdezésének eredményét látjuk.

```

<!-- HVG ./0116/0116009.htm -->
<div type="article" column="unspec">
<opener>
<dateline>
<w lemma="HVG" msd="N.NOM" ctag="NS3NN">HVG</w>
<w lemma="2001/16" msd="DIG" ctag="Q">2001/16</w>
<c lemma="." msd="SPUNCT" ctag="SPUNCT">.</c>
<w lemma="szám" msd="N.NOM" ctag="NS3NN">szám</w>
<date iso8601="04-21-2001">
<w lemma="2001._április_21." msd="DATUM"
ctag="DATUM">2001._április_21.</w>
</date>
</dateline>
</opener>
<head rend="IT" type="unspec">
<s>
<w lemma="egészségügyi" msd="A.NOM" ctag="AS_A">Egészségügyi</w>
<w lemma="szigorítás" msd="N.PL.NOM" ctag="NP3NN">szigorítások</w>
</s>
</head>
<head>
<s>
<w lemma="sok" msd="Num.NOM" ctag="Q">Sok</w>
<w lemma="zsebből" msd="N.ELA" ctag="NS3NE">zsebből</w>
<w lemma="vérzik" msd="V.e3" ctag="VS3RI">vérzik</w> </s>
</head>
<head rend="BO" type="display">
<s>
<w lemma="Alaposan" msd="Adv" ctag="R">Alaposan</w>
<w lemma="felkavar" msd="Pre.V.TMe3" ctag="@VS3PD">felkavarta</w>
<w lemma="a" msd="Det" ctag="D">a</w>
  
```

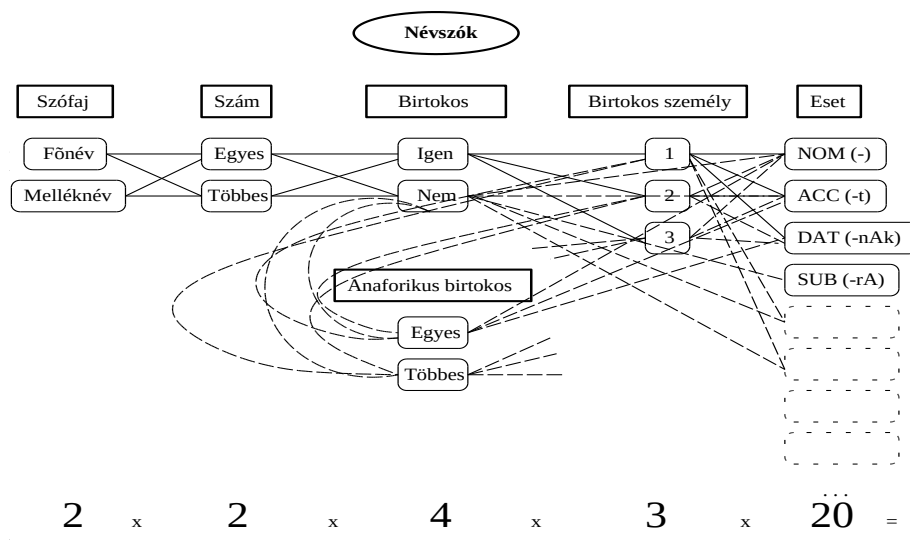
3. ábra A MNSZ kódolása



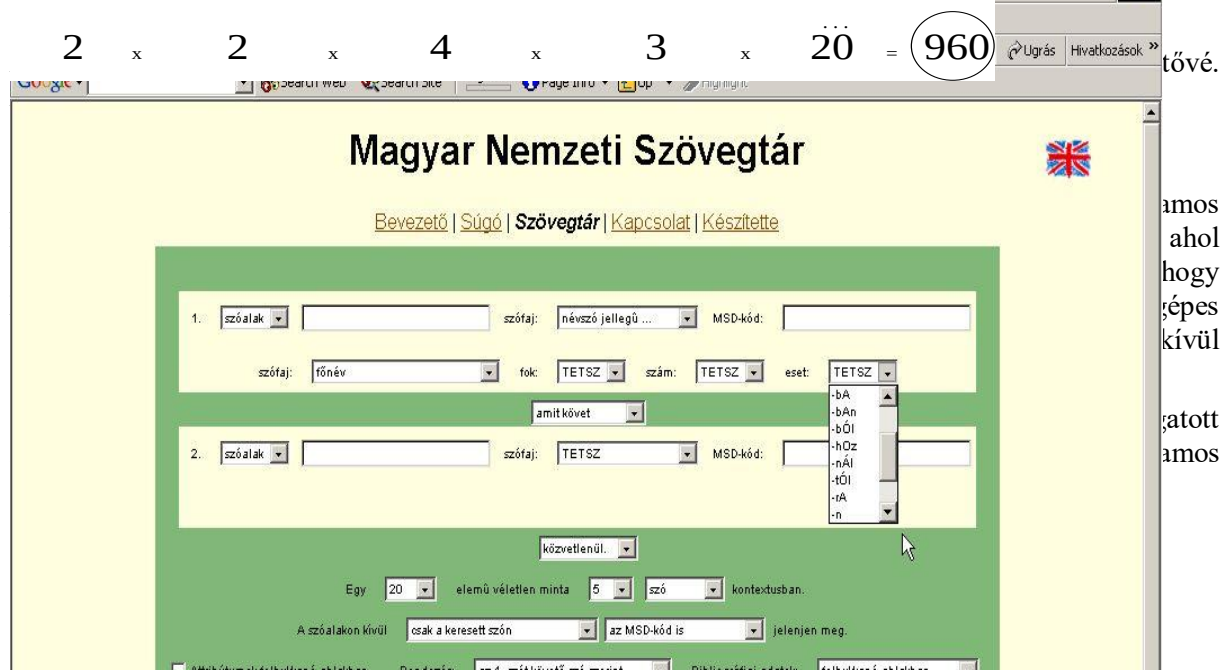
4. ábra A "bokor" szó alakváltozatainak lekérdezése az MNSZ-ben

Néha sem egy szótó, sem egy szóalak megadása nem vezet kellő eredményre. Az MNSZ-ben 656272 helyen fordul elő az *az* szó, ami egyaránt lehet határozott névelő és mutató névmás. Ha tehát a mutató névmási használatot akarjuk vizsgálni, ennyi alakot kellene átvizsgálni, hogy megállapítsuk, melyik esetről van szó. Szerencsére azonban a nyelvtani jellemzők felhasználásával pontosíthatjuk kérésünket, és így könnyedén megtalálhatjuk a keresett mutató névmási előfordulásokat.

A Magyar Nemzeti Szövegtár minden egyes szövegszóról kimerítő nyelvtani elemzést tárol a morfoszintaktikai jellemzőkben. A magyar nyelv rendkívül gazdag alaktani rendszere miatt ez a névszók esetében majdnem ezer variációt jelent (ld. 5 ábra).

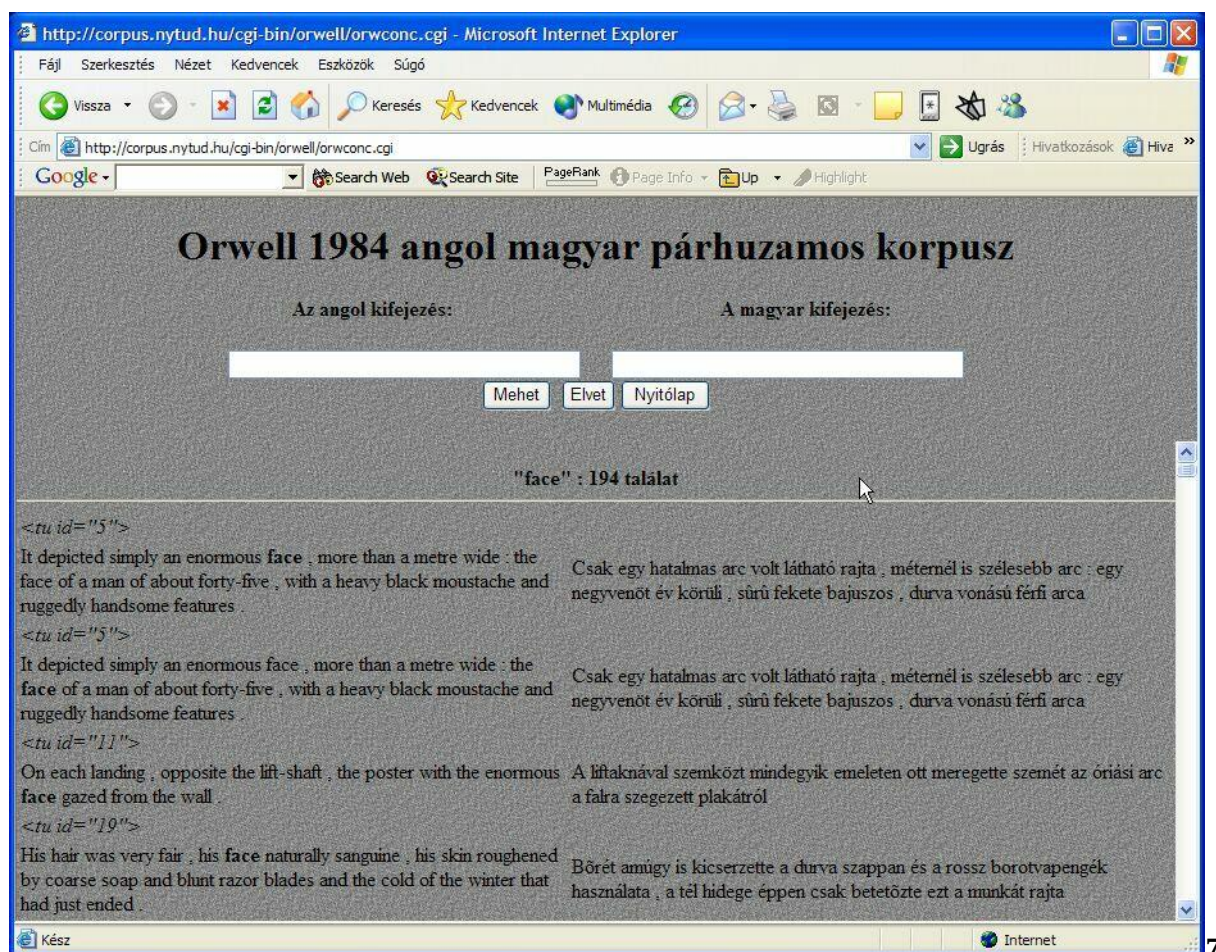


y miképpen lehet a nvalóan nem lehet nellyel a nyelvtani alkotnak, amelyek zható igékre, ezért íst tehát egy olyan i. Erre mutat példát



korpusz, amely a TELRI egyesület által terjesztett CD-ROM-on megtalálható (Erjavec et al. 1998). Ez a korpusz George Orwell 1984 című regényének angol eredetijét és annak magyar fordítását tartalmazza. A korpusz értékét növeli, hogy akárcsak az MNSZ, ez is gondosan van nyelvileg annotálva: minden szövegszó morfológiai elemezve és egyértelműsítve van. A párhuzamos korpuszok vizsgálatának egyik eszköze a párhuzamos konkordancia. Ez olyan szoftver, amely bemutatja valamely keresett szó összes előfordulását a forrásnyelvben, és minden előfordulás mellett bemutatja azt az célnyelvi egységet, amelyre a két korpusz illesztve van. Ez általában a mondat, így van ez az Orwell-korpusz esetében is. Tehát a párhuzamos konkordancia egyik oldalán az a forrásnyelvi mondat szerepel, melyben előfordul a keresett szó vagy kifejezés, a másik oldalán pedig az a mondat, amely ennek a forrásnyelvi mondatnak a fordítása. Ilyen konkordancia segítségével vizsgálható az Orwell-korpusz a <http://corpus.nyud.hu/orwell> webcímen.

Amint azt a 7. ábrán láthatjuk, a lekérdező rendszernek nemcsak a keresett kifejezést adhatjuk meg, a bal oldali rovatban, hanem a mellette lévő kis ablakban azt is meghatározhatjuk, hogy mely fordítási kifejezést keresünk a célnyelvi oldalon. Ha csak a forrásnyelvi kifejezést adjuk meg, akkor valamennyi olyan mondatot, visszakapunk, amely a forrásnyelvi mondat fordítása, ha viszont a kívánt fordítási megfelelést is megadjuk a jobb oldali ablakban, akkor csak azokat a mondatokat kapjuk vissza, amelyek nemcsak a forrásnyelvi kifejezést tartalmazzák, de a nekik megfelelő célnyelvi mondatban előfordul a megadott célnyelvi kifejezés is.



. ábra Az Orwell 1984 angol-magyar párhuzamos korpusz lekérdezése

Lexikai adatbázisok

A Nyelvtudományi Intézet Korpusznyelvészeti osztályán folyó munkálatok során nemcsak szöveges, hanem lexikai adatbázisok is készültek. Ezek között három munkát ismertetünk röviden.

Az Értelmező Kéziszótár

Az Értelmező Kéziszótár jelenleg készülő új változata kapcsán felmerült az igény, hogy az már korszerű, SGML kódolt elektronikus változatban szülessen meg. Az ezzel kapcsolatos elméleti és módszertani munkát a CONCEDE (CONsortium for Central European Dictionary Encoding) projektum keretében végezte el a Nyelvtudományi Intézet Korpusznyelvészeti osztálya. A projektum célja az volt, hogy a résztvevő hat kelet-európai nyelv (cseh, magyar, szlovén, román, bolgár, észt) lexikográfiai sajátosságait figyelembe véve kidolgozzon egy egységes notációt és technológiát e nyelvek korszerű elektronikus szótárai számára. Köztudomású, hogy a szótárak szerkezete rendkívül összetett. A cél olyan kódolási eljárás elkészítése volt, amely nemcsak számítógéppel olvasható, hanem a számítógép számára értelmezhető egyértelműséggel és explicitiséggel tartalmazza a szótári egységek leírását. Ennek kidolgozott szótárkódolási technológia az ÉrtSz. 2500 szócikkén meg is valósult, melynek eredményeként létrejött mintegy 8000, a concede.dtd-nek megfelelően szabványosan kódolt szócikk (Váradí–Oravecz 1999).

A Magyar Ragozási Szótár

Elekfi László munkája (Elekfi 1994) eredetileg az Értelmező Kéziszótár számára készült ragozási útmutatóból lett egy önálló, az ÉrtSz. teljes szócikkállományát feldolgozó szótár. A rendkívül gondos, aprólékos munka a legrészletesebb, legjobban kifejtett tárgyalása a magyar ragozási rendszernek. Bármennyire is explicit és körülhatárolt a mű, mégis könyv alakjában közzétett változata csak emberi tudás segítségével értelmezhető. Nyilvánvaló, hogy a szótár nem tartalmazhatja az összes lehetséges toldalékolt szóalak tételes felsorolását, hiszen a magyar morfológia gazdagsága egyszerűen kezelhetetlen méretű könyvet eredményezne. Ehelyett paradigmatablákba foglalva adja meg a szavak ragozását. Még a paradigmatablák is hierarchikus rendszert alkotnak, és általában egy-egy szó teljes ragozását két lépcsőben kapjuk meg.

Ahhoz, hogy a Ragozási szótár gazdag tartalmát számítógép számára nemcsak olvasható, de értelmezhető alakra hozzuk, a szótárban rejlő implicit információt explicit alakra kell hoznunk. Ennek első lépéseként minden egyes paradigmatablát elő kellett állítani, azaz öröklött jegyeit az adott paradigma egyéni jegyeivel együtt legenerálni a főalakváltozatok pontos feltüntetésével. További feladat volt a toldalékkombinációk előállítás, melyre a szótár eredetileg nem tér ki. Ez lényegesen megnövelte az egyes paradigmatablák méretét: egy eredetileg 36 tételtől álló melléknévi paradigma a toldalékkombinációk kifejtése után 2027 alakot tartalmaz. Végül a morfológiai adatbázissá alakítás megkívánta azt is, hogy a szótár eredeti célját meghaladva a képzőket is bevonjuk az alakváltozatokba.

Vonzatszótár

A morfológiai adatbázis elkészítése után a Nyelvtudományi Intézet Korpusznyelvészeti Osztályán 2001 nyarán elindult egy olyan munka, amely a lexikográfia és a szintaxis határmezsgyéjén húzódik. Célja egy adatbázis létrehozása, amely minden olyan nyelvi információt tartalmaz, amely szükséges lehet a magyar nyelv szintaktikai elemzéséhez, és az egyedi lexikai elemekhez kapcsolódik, azaz nem szóosztályokon működő szintaktikai szabály. Szintaktikai elemzés alatt természetesen számítógéppel végzett elemzést értünk, ami az adatbázisba foglalt nyelvi információk minden eddiginél explicitebb kifejtését kívánja meg. Az adatbázis szóanyagát a Magyar Nemzeti Szövegtár leggyakoribb 20 ezer szava alkotja. Központi részét az igék adatbázisa alkotja, amely több mint háromezer igei lemmából áll. Az igei adatbázis jelenti az első kísérletet arra, hogy létrehozzuk az igék vonzatkeretét. Az alany és tárgy mellett *vonzat1* és *vonzat2* néven két további vonzatot kódolunk. A vonzatokat felszíni esetvégződésük szerint, a tematikus szerep megjelölése nélkül tartjuk számon. Az eset mellett az adatbázis tartalmaz megszorító szabályokat is, melyek a mondat főbb összetevőinek (alany, tárgy) jegyeire hivatkoznak (pl. élő alany, absztrakt tárgy stb.) A vonzatkeret mellett feltüntettük a főmondat és az ige komplemenseként szereplő tagmondat közötti koreferenciális viszonyokat is. A főnevek szemantikai kódolásakor a szokásos *+élő*, *+absztrakt*, *+human* stb. mellett olyan jegyeket is alkalmaztunk, mint a *+testrész*, *+mérték*, *+időegység*, *+időjárás*.

Hivatkozások

Alexin et al. (2001)

Alexin, Zoltán – Péter Leipold – János Csirik – Károly Bibok – Tibor Gyimóthy. A Rule-Based Tagger Development Framework. In: Popelínský, Lubos – Nepil, Miloslav (eds.) *Proceedings of the 3rd Workshop on Learning Language in Logic*, 1–10. Strasbourg, France, 2001.

Christ (1994)

Christ, Oliver. A Modular and Flexible Architecture for an Integrated Corpus Query System. In: *Papers in Computational Lexicography (COMPLEX-94)*, 23–32. Linguistics Institute, Budapest, 1994.

Elekfi (1994)

Elekfi László. *Magyar ragozási szótár*. MTA Nyelvtudományi Intézet, Budapest, 1994.

Kiss et al. (2001)

Kiss, Gabriella – Kiss, Margit – Pajzs, Júlia. Normalisation of Hungarian Archaic Texts. In: *Papers in Computational Lexicography (COMPLEX-01)*. University of Birmingham, Birmingham, 2001.

Oravecz – Dienes (2002)

Csaba Oravecz and Péter Dienes. Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002*, 710–717. Las Palmas, Spain, 2002

Pajzs–Váradi (1997)

Pajzs Júlia – Váradi Tamás. A magyar irodalmi és köznyelv nagyszótárának korpusza a HUNGARNET közösség számára. *Workshop-97 konferencia*. NIIF, Budapest, 1997.

Prószéky–Tihanyi (1996)

Prószéky, Gábor – Tihanyi, László. HUMOR – a morphological system for corpus analysis. *Proceedings of the 1st TELRI Seminar in Tihany, Budapest*, 149–158.

Váradi (1997)

Váradi, T. 1997. Adatok a nagyszótári korpusz összetételéről in Kiss G. (szerk.) *Szavak – nevek – helyek. Kiss Lajos 75. születésnapjára*. Budapest: MTA Nyelvtudományi Intézet, 437–446.

Váradi (1999)

Váradi Tamás. On Developing the Hungarian National Corpus. In: Vintar, S. (ed.) *Proceedings of the Workshop Language Technologies – Multilingual Aspects. 32th Annual Meeting of the Societas Linguistica Europea*, 57–63. Ljubljana, Slovenia, 1999.

Váradi (2000)

Váradi Tamás. A nyelvhasználat empirikus vizsgálatáról. In: Andor József, Szűts Tibor, Terts István (szerk.) *Színes eszmék nem alszanak... (Szépe György 70. születésnapjára)*, Lingua Franca, Pécs, 1285–1291, 2001.

Váradi – Oravecz (1999)

Váradi, Tamás – Csaba Oravecz.. Morpho-syntactic Ambiguity and Tagset Design for Hungarian. In: *Proceedings of the EACL LINC Workshop on Annotated Corpora*. Bergen, Norway, 1999.