A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai

Gulyás Attila[1], Galántai Júlia[1], Szabó Martina Katalin[1, 2, 3], Szebeni Zea[1]

[1]MTA TK „Lendület" RECENS Kutatócsoport
[2]Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3]Precognox Informatikai Kft.
{Gulyas.Attila;Galantai.Julia;Szabo.Martina;Szebeni.Zea}@tk.mta.hu

Kivonat: Jelen dolgozatban egy magyar nyelvű, spontán beszélt nyelvi korpusz, a HuTongue leiratozásának és annotálásának minőségbiztosítási munkálatairól számolunk be. A korpuszban feldolgozott szövegeket hétköznapi szituációkban, külső ingerektől teljesen elzárt környezetben keletkezett spontán nyelvi produktumok alkotják. A korpusz létrehozásának legfőbb célja, hogy megfelelő vizsgálati anyagot teremtsünk a pletyka természetének elsősorban társadalomtudományi szempontú kutatásához. A HuTongue egy egyedülálló adatbázis: tudomásunk szerint ez az egyetlen magyar nyelvű, nagy méretű, spontán szituációkban keletkezett, beszélt nyelvi korpusz, amely teljes egészében manuálisan gépelt és annotált formájú. A korpusz létrehozása – amely jelenleg is folyamatban van − több munkafázisban történik. Az előkészítés után a fájlokat egy feldolgozócsapat legépeli és annotáltatja. A munka három alapvető feladatból tevődik össze: a hanganyagon hallható verbális közlések rögzítéséből, a nem verbális hanghatások kódolásából, valamint egy, a kutatás szempontjából kardinális, szemantikai-−pragmatikai jellegű sajátság jelöléséből. Azt reméljük, hogy a korpusz a kutatási kérdés sokrétű és automatikus megoldásokkal hatékonyan támogatott vizsgálatát fogja lehetővé tenni a számunkra a jövőben. A jelen dolgozat célja, hogy bemutassuk e komplex feldolgozási munkának a minőségbiztosítási folyamatát. Szólunk a minőségbiztosítás szempontjairól, megtervezésének dilemmáiról és lépéseiről, valamint bemutatjuk az általunk alkalmazott megoldást.

1. Bevezetés

A magyar nyelvű, spontán beszélt nyelvi korpusz, a HuTongue korpusz létrehozásának legfőbb célja, hogy megfelelő vizsgálati anyagot teremtsünk a pletyka természetének elsősorban társadalomtudományi szempontú, beható vizsgálatához.

Annak ellenére, hogy az elmúlt évtizedekben számos spontán nyelvi korpusz született a világ különböző nyelvein [4], [1], [7], [6], a magyar nyelvre irányuló kutatásunkhoz nem állt rendelkezésünkre megfelelő vizsgálati adatbázis. Mindenekelőtt, a magyar nyelv legtöbb beszélt nyelvi korpusza olvasott szövegekből áll [3]. Jelenleg három hazai spontán beszélt nyelvi korpuszról van tudomásunk, az ún. BEA (Magyar spontán

beszéd adatbázis) [3], a Kivi (Korpusz az inferencialitás vizsgálatához) [5], valamint egy készülőben lévő adatbázisról, a Budapesti Egyetemi Kollégiumi Korpuszról (BEKK, bekk.elte.hu), amely társadalomtudományi céllal készül és nyelvi interakciókat tartalmaz. Ugyanakkor, saját vizsgálati céljainknak közülük egyik adatbázis sem felelt meg részletesebben [2], [8]. Olyan korpuszra volt szükségünk, amely megfelelő ahhoz, hogy spontán nyelvi környezetben keletkezett diskurzusokat vizsgálhassunk, és a pletyka jelenségére vonatkozó megállapításokat tehessünk. A fentebbieket megfontolva döntöttünk egy saját korpusz létrehozása mellett. A korpusz létrehozásának fő célja az volt, hogy mind a szövegek típusa, mind az alkalmazott annotáció tekintetében olyan korpuszt hozzunk létre, amely lehetővé teszi a pletyka jelenségének beható vizsgálatát.

Egy korábbi dolgozatunkban [8] részletesen tárgyaltuk a korpusz szövegeinek keletkezési körülményeit, a feldolgozói munka előkészítő lépéseit, a feldolgozási és az annotálási folyamat eszközét, módszereit, a megtervezés dilemmáit, valamint az alkalmazott megoldásokat. A jelen dolgozatban a korpusz leiratozásának és annotálásának minőségbiztosítási folyamatát kívánjuk ismertetni. Meg szeretnénk mutatni mindazokat a minőségbiztosítást érintő kérdéseket és dilemmákat, amelyekkel a munka során szembesültünk, valamint a feladatban alkalmazott megoldásokat és eszközöket. A tárgyalás során, a problémák szemléltetése céljából a korpuszból származó példákat is segítségül hívunk.

2. A korpusz rövid bemutatása és alapvető statisztikai adatai

A korpuszban feldolgozott szövegeket hétköznapi szituációkban, külső ingerektől teljesen elzárt környezetben keletkezett spontán nyelvi produktumok alkotják. A spontán alatt azt értjük, hogy a résztvevők azzal és arról beszéltek, akivel és amiről akartak. Emellett a beszéd mennyisége sem volt korlátozva. A környezet, amelyben a rögzítés történt, egy összesen három nagyobb társas térre osztható épület volt. A résztvevők átlagos életkora 23 volt, a legfiatalabb 21, a legidősebb pedig 26. Közülük 3 nő és 5 férfi volt. A hangrögzítés napi 24 órában történt úgy, hogy minden résztvevő rögzítő eszközt viselt.[1] A résztvevők tudatában voltak annak, hogy a hangjukat 24 órában rögzítik, és a hanganyag jogtulajdonosa hozzájárult annak kutatási célú felhasználásához.

A teljes anyagból végül nyolc nap felvételét dolgoztuk fel a korpuszban. Ez az anyag összesen megközelítőleg 500 órányi hanganyagot tesz ki.

A szövegek feldolgozásának a megkezdése előtt az anyagot a munkához elő kellett készíteni: a felvételekből a csendet, és az egyéb, nem beszéd tartalmú hangokat tartalmazó részeket a lehető legteljesebb mértékben ki kellett vágnunk, majd az anyagot 60 perces egységekre osztottuk [8].

[1] A spontán nyelvi hanganyagot, amelyet egy szórakoztatóipari cég rögzített, kizárólag tudományos célokra adták át és használjuk fel, teljes titoktartási kötelezettségvállalás mellett. A felvételen résztvevő önkéntesek teljes körű tájékoztatásban részesültek a hangfelvételek elkészüléséről.

3. A feldolgozói munka rövid bemutatása

A feldolgozói munkát a résztvevők kiválasztása, majd betanítása előzte meg. A kiválasztott feldolgozók először egy nagyon részletes ún. útmutatót kaptak, amely pontos leírását adta az elvégzendő feladatnak, valamint a feldolgozáshoz használni kívánt eszköz kezelésének. Az útmutatóban rendre példákat (szöveges és hangzó) is elhelyeztünk. A leírás megismerése után a teljes folyamatot megbeszéltük a feldolgozókkal.

Annak céljából, hogy a munka minőségét a lehető legmagasabb szinten tarthassuk, a feldolgozókkal folyamatos kapcsolatot tartunk, a kérdéseket, észrevételeket megbeszéljük, és a megbeszélteket a teljes csoport számára elérhető formában rögzítjük.

A feldolgozáshoz az f4transkript (https://www.audiotranskription.de/english/f4.htm) elnevezésű szoftvert alkalmaztuk. A szoftvert gyakran alkalmazzák társadalomtudományi kutatások alkalmával, mert a használata nagyban elősegíti a leiratozás és a taggelés gyors és egyszerű, egy időben történő elvégzését. Mivel a jelölés során nem fonetikai annotációt készítettünk, a számunkra kiválóan megfelelt a szemantikai és pragmatikai jellemzők rögzítésére.

A feldolgozók online kapták meg a hangfájlokat, valamint az egyes hangfájlokhoz tartozó szegmensek időbélyegeit - amelyek a kivágott részek jelölésére szolgálnak, és útmutatóként a fájlok összeillesztéséhez, txt formátumban. Ez utóbbi tette lehetővé számunkra azt, hogy a későbbiekben az elkészült, írásos formájú szegmenseket az időbélyegek alapján egymáshoz illesszük, valamint útmutatóként szolgálnak a kivágott szöveganyag jelzésére. A feldolgozók e fájlt töltötték ki a program segítségével a hanganyag írásos rögzítésével. A alábbi ábra egy részletet mutat a program működéséről.

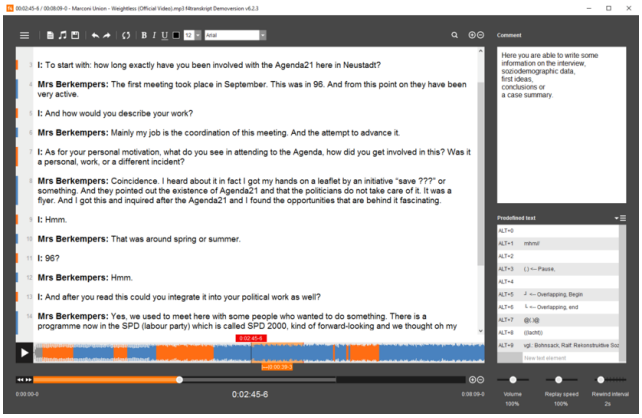

Fig. 1. Részlet az F4-programról

A feldolgozók munkája a jelen fázisban három alapvető feladatból tevődött össze: a hanganyagon hallható verbális közlések rögzítéséből, a nem verbális hanghatások kódolásából, valamint egy, a kutatás szempontjából kardinális, szemantikai--pragmatikai jellegű sajátság jelöléséből. A munka során tehát nem csupán a verbális információk rögzítését céloztuk, hanem olyan, nem verbális információk tagelését is, amelyek véleményünk szerint fontos segítségül szolgálhatnak majd számunkra a korpusz

jövőbeli felhasználása során a kutatási probléma mélyebb összefüggéseinek feltárásában [8]. A nem verbális hanghatások tageinek kiválasztásánál és meghatározásánál fontos szempont volt, hogy a kutatási kérdéseink és hipotéziseink megválaszolásához megfelelő elemzési szempontokat és magyarázó változókat szolgáltassunk. Az annotációs jelölések kiválasztásánál tehát fontos volt az olyan érzelmi töltetű megnyilvánulások jelölése, amelyek korrelációja az általunk jelölt pragmatikai sajátságok meghatározásánál fontos többlettartalmat jelölhet (pl. gunyoros vagy zavarodott nevetés, köhögés stb.)

Általános, a teljes leiratozásra vonatkozó sajátságként mondható el, hogy a feldolgozóknak a szoftver segítségével időbélyegeket kell elhelyezniük a gépelt szövegben, annak megfelelő helyein (részletesebben l. [8]). Ez nem csak a fájlok összeillesztését teszi lehetővé, hanem a későbbiekben időintervallum alapú mérési eszközök kidolgozására ad lehetőséget, és megőrzi a hanganyag kapcsolatát a szöveges fájllal.

(1) #00:00:46-0# (Sanyi) Persze. #00:00:46-4#
#00:00:46-4# (Gabi) Sietünk, hátha #00:00:49-3#

Azt, hogy az adott diskurzus verbális és nem verbális hanghatásai kitől származnak, ugyancsak jelöljük, például

(2) (Gabi) Hol a kávém?

Tekintettel arra, hogy minden egyes résztvevő mikroportjának hanganyagát külön feldolgoztuk, nem szükséges az, hogy az anyagok teljes tartalmát leiratoztassuk, elég csupán azt, ahol a mikroport viselője további szereplőkkel részt vesz egy diskurzusban. Minden diskurzusnál jelöljük azonban, hogy a verbális és előre meghatározott nem verbális hanghatások kitől származnak.

Amennyiben érthetetlen a teljes megszólalás vagy annak egy része, a következő jelet alkalmazzuk a diskurzus megfelelő helyén: (?). Abban az esetben pedig, ha a leiratozó nem biztos benne, hogy jól értette, amit hallott, az adott szövegrészt a következő nyitó- és zárótaggel jelöli meg: (( )) Ha olyan megszólalót hall, aki nem tartozik a csoport tagjai közé, így jelöli: (k?)

Két típusú hanghatást kódolunk, a pillanatnyit, valamint a hosszabb ideig tartót. Az előbbiek közé tartozik például a köhögés, a nevetés vagy az ásítás. A hosszabb ideig tartó hanghatások közé tartozik például az, ha valaki sírva vagy nevetve mond valamit. Ilyenkor nyitó- és zárótaget használunk, annak érdekében, hogy beazonosíthassuk, milyen hosszú egy-egy hanghatás. Az alábbi példák egy nevetést, valamint egy nevetve mondott szövegrészt tartalmaznak:

(3) a. (Gabi) Szerintem (~) elég jó csaj!
b. (Gabi) Szerintem (~) elég jó csaj! (~)))

Amennyiben több hanghatás is történik egyszerre, akkor azt a megfelelő sorrendben és feltétlenül ugyanabban a sorban, azaz ugyanazzal az időbélyeggel jelöljük.

(5) #23:18# (Gabi) Mikor lesz kész a vacsora? (Éva) (~) #23:33#

Bár nem jegyezzük le a verbális és a nem verbális hanghatásokat azokban az esetekben, amikor azok nem az adott diskurzus részét képezik, azonban bizonyos, a kutatás szempontjából releváns információkat ilyenkor is jelölni kell. Ezekben az esetekben három különböző megoldást alkalmazunk annak a jelölésére, hogy mit érzékelünk a háttérben levőkről [8]. Amennyiben kivehetők a nem az adott diskurzus résztvevőinek a verbális és nem verbális közlései, és be is tudjuk azonosítani a beszélőket, akkor a résztvevők neveinek a kezdőbetűivel jelöljük a jelenlétüket, például Gabi esetében (G) taget alkalmazunk. Ha, bár a közlések kivehetők, a forrásukat azonban már nem tudjuk beazonosítani, akkor megpróbáljuk megbecsülni a számukat, például (4). Amennyiben még ez sem lehetséges, a következő annotációs jelet alkalmazzuk: (t?).

A leiratozási munka egy, a kutatás szempontjából kardinális lépése az, hogy a verbális közlések meghatározott tartalmait (p) jelzéssel látjuk el. A (p)-tartalmat a következőképpen definiáljuk: (p) jelzéssel jelöljük azokat a szövegrészeket, amelyek során a beszélgetésben egy olyan, a csoporthoz korábban vagy jelenleg is tartozó harmadik személyről esik szó, aki valószínűleg nincs jelen a diskurzus közben. Amennyiben tudjuk, hogy az adott közlés kire vonatkozik, azt is jelölni kell a megfelelő módon.

(6) (Gabi) jól alszik Éva a másik szobában munka helyett! (p-É)

4. A minőségbiztosítás szempontjai

Az 1. táblázat tartalmazza a jelen munka szempontjából fontos kódokat, és azok, a minőségbiztosítás szempontjából lényeges jellemzőit.

A táblázatban szereplő *szegmens* terminus alatt a két időbélyeg közötti szövegrészeket értjük, melyek több mondatból állhatnak és több jelölőt is tartalmazhatnak. A különböző jelölők közül a verbális tartalmaknál, valamint az azokhoz rendelt jelölők (hanghatások) esetében a szegmensen belüli elhelyezkedés is fontos, a többi jelölő esetén csupán az a lényeges, hogy jelen vannak-e, vagy sem.

A gépelt kimenettől, valamint a manuális annotációtól azt reméljük, hogy mind a társadalomtudományi, mind a nyelvészeti kutatási kérdéseink megválaszolásában lényegi segítséget nyújthatnak. Ehhez azonban elengedhetetlen egyrészt az, hogy az elvégzett munka belső érvényessége megfelelő legyen, másrészt, hogy a hibás gépelés vagy annotálás ne okozhasson fennakadást az elemzés során. A feldolgozók produktumait illetően tehát komoly minőségbiztosításra van szükség a folyamat lefolytatásához.

1. Táblázat: A jelen munka keretében ellenőrzött kódok, és azok lényegi sajátságai

Jelölő alkalmazása	Jelölő	Kategória	Leírás	Pozíció fontos
Szegmens eleje	(<név>)	Beszélő	A beszélő keresztneve	Nem
Szegmens-részlet	(<hanghatás>)	Hanghatás	<hanghatás> lehet: "s" - sóhajtás, "~" - nevetés, "*" - sírás, "gn" - gunyoros nevetés, "zn" - zavarodott nevetés	Igen

Szó	(<hang-hatás>)	Hang-hatás	<hanghatás> lehet: "sik" - sikítás, "k" - köhögés, "á" - ásítás, "pi" - pisszegés, "ujj" - ujjongás, "tor" - torokköszörülés, "f" - fütyülés, "é" - éneklés	Igen
Szegmens	(t?)	Tisztázó	távolabbi beszélők jelenléte	Nem
Szegmens vége	(k?)	Tisztázó	nem szereplő beszéde	Nem

5. A minőségbiztosítás lépései és eredményei

Tekintettel a feldolgozói munka komplexitására, egy alaposan átgondolt és részletes minőségbiztosítási folyamatot kellett megterveznünk a leiratozók munkaminőségének ellenőrzéséhez, illetve a minőség folyamatos biztosításához.

A HuTongue korpusz létrehozásának a folyamata során több minőségbiztosítási módszert is alkalmaztunk. Első lépésben a feldolgozott fájlok szúrópróba-szerű, kvalitatív átnézését végeztük minden egyes feldolgozó munkájának az ellenőrzése céljából. A szúrópróba-szerű ellenőrzés lehetőséget nyújtott ahhoz, hogy egyes feldolgozók szisztematikus hibáit azonosítsuk, és visszajelezzünk arról az egyes feldolgozók számára. Az ilyen módon feltárt hibákat, hiányosságokat a feldolgozók javították, azonban túl nagy minőségbeli eltérés esetén, abban az esetben, ha az első visszajelzés sem hozott eredményt a feldolgozó munkájának javulása terén a feldolgozó nem folytathatta a korpusz feldolgozását. Az ilyen típusú, rontott fájlokat később újra feldolgoztattuk egy másik feldolgozó munkatárs segítségével.

A korpusz minőségbiztosítása érdekében azonban szükségünk volt egy olyan átfogó mérőeszközre, amely több dimenzióban is képes mérni a munka létrehozásának eredményességét, pontosságát. Ezért a feldolgozók munkáját egymáshoz viszonyítva, és egy referenciafeldolgozó munkájához képest is ellenőriztük. A referenciagépelő kiválasztását a leiratozott fájlok elsődleges, kézi ellenőrzése alapján végeztük. Így a kézi ellenőrzés során, a konzisztensen legjobban teljesítő feldolgozót választottuk ki a kvantitatív ellenőrzés referenciagépelőjének.

A kvantitatív minőségbiztosítási folyamatot ezért a következő részfeladatokra bontottuk: Szövegegyezés, annotáció egyezése és az időbélyegek elhelyezésének helyessége.

A továbbiakban bemutatjuk a minőségbiztosításunk fókuszában álló jellemzőket, az erre kidolgozott mutatót, azaz az IRI indexet (Intercoder Reliability Index), valamint a mutató által szolgáltatott eredményeket egy 60 percet átfogó hangfájl részletén.

5.1 A mérni kívánt sajátságok

A HuTongue korpusz egyik legkülönlegesebb jellemzője az általános nyelvi korpuszokhoz képest a benne alkalmazott komplex és széles körű annotáció. Ebbe beletartozik a korábbiakban ismertetett, több dimenziós annotáció, illetve a szövegek időbélyegekkel történő kiegészítése. A minőségbiztosítás célja tehát a következő

sajátságok pontosságának a mérése volt: a gépelt szöveg helyessége, a jelölők, valamint az időbélyegek megfelelő használata.

A szövegek helyessége jelen kutatásunkban nem a magyar helyesírás szabályainak való megfelelést jelenti, hanem − élőnyelvi korpusz lévén − az elhangzottak lehető legpontosabb rögzítését, bizonyos esetekben fonetikus átírással. Ennek megfelelően a legépelt szöveg nyelvtani helyessége számunkra nem befolyásolja a kapott eredményeket.

A tagek tartalmazzák kutatásunk szempontjából a legfontosabb információt, ezért ezekre helyeztük a legnagyobb hangsúlyt a minőségbiztosítás során. A legfontosabb szempont a különböző jelek egy szegmensen belüli megléte vagy hiánya volt, valamint néhány jelölő esetében fontos volt azok mondaton belüli elhelyezkedése is. A legtöbb tag esetében a meglét ellenőrzése elégséges, azonban, különösen a hanghatásokat jelző jelölők esetében a helyes elhelyezés is fontos szempont. További nehézséget jelentett a nyitó-, illetve zárótagek jelölésének mérése is egyes kódok esetében, ezeknek a jelölőknek a pontos bemérése, összehasonlítása a szöveg egyes részein.

Végül, az időbélyegek használatát illetően mind a pontosság, mind a meglét ellenőrzése fontos volt. A feldolgozók egzakt kritériumok szerint tehetik ki az időbélyegeket a sorokvégén (a szegmenseket természetesen mindig időbélyeg zárja), így ezek konzisztens használata is nagyon fontos. Az időbélyegek adják a szöveg természetes szegmentálását, ezért ezek pontossága nagy jelentőséggel bír. Az, hogy egy adott feldolgozó mekkora szegmenseket alkot, tehát milyen gyakran tesz ki egy időbélyeget, praktikus szempontból fontos számunkra, de követnie kell a beszéd természetes meghatározottságát is, ezért a nem megfelelő helyen való tagolást is hibának tekintettük.

A fentebb ismertetett három szempont együttesen adja meg azt a feltételrendszert, amely mentén egy adott feldolgozó munkája kiértékelhető. Fontos azonban, hogy az értékelés ne csupán kvalitatív módon (azaz humán ellenőrzéssel), hanem kvantifikálhatóan is megtörténhessen. A minőségbiztosítást ugyanis iteratív folyamatként kell értelmeznünk, tehát meghatározott periódusonként ismételt minőségbiztosítási fázisokat kell beiktatnunk a feldolgozói munkába, amely feladat manuálisan egyrészt nem idő és költséghatékony, másrészt a kvantitatív értékelésre sem ad lehetőséget.

5.2 A megvalósítás eszköze és módszere

Az eddigiekben ismertetett három szempontot egyaránt lehetséges külön-külön mérőszámokkal és kompozit mérőszámokkal is mérni. A jelenlegi munkánk során egy kompozit mérőszám előállítására törekedtünk, amelynek segítségével általános képet kaphatunk a feldolgozók munkájáról, azonban, amennyiben ez szükséges, a három sajátság eredményességét külön-külön is fel tudjuk mérni minőség javítási céllal.

5.2.1 Az IRI index

A minőségbiztosítás során az egyik legfontosabb szempont volt az alkalmazott mérőszám(ok) egyszerű értelmezhetősége, illetve egyúttal a részletekbe menő információgyűjtés. Ahhoz, hogy mindezt egyben elérhessük, definiáltuk az Intercoder Reliability Indexet, melyre a továbbiakban az IRI rövidítéssel hivatkozunk. Ez a mérőszám két szöveg összehasonlítására alkalmas, amelyben a T2 a referenciaszöveget, a

T1 pedig a kiértékelt szöveget jelöli. A lehetséges értékei 0 és egy között mozognak, ahol 1 jelöli a referenciaszöveggel történő teljes egyezést (azaz, hogy a feldolgozó munkája megbízható), 0 pedig az abszolút különbözőséget (azaz, hogy a feldolgozó munkája nem megbízható). Az IRI-t az alábbi összefüggéssel definiáljuk, amelyet a három lényeges minőségbiztosítási szempont mérőszámaiból összeállított kompozit indexként értelmezünk:

$$IRI = 1 - \left[w_L L(T_1, T_2) + w_{Ta} I_{Ta}(T_1, T_2) + w_{Ti} I_{Ti}(T_1, T_2) \right] \quad (7)$$

A kifejezés első eleme $L(T_1, T_2)$ a szövegek közti eltéréseket magába foglaló mutató, amely célra a legmegfelelőbbnek az egyszerű relatív Levenshtein távolságot találtuk[2]. Ez a mutató nem veszi figyelembe a helyesírás szabályait, és kifejezetten a szövegek közti eltérés mértékét adja meg, tehát a számunkra lényeges információt mutatja.

A szövegek közti eltérések mutatóját a jelölők pontosságát leíró mérőszám követi, melyet $I_{Ta}(T_1, T_2)$-vel jelölünk. Ez már önmagában egy kompozit mérőszám, amelyeket jelölőkategóriánként is számolunk, majd összegzünk. A mérőszám magában foglal egy a jelölők meglétéből és elhelyezkedéséből számolt pontszámot, az alábbi módon: egy adott jelölő hiánya (akár a referenciaszövegben, akár a kiértékelt szövegben) 1 pont, egy adott jelölő kategórián belüli eltérése 0,5 pont, továbbá − ha az adott kategória esetén ez releváns − egy jelölő rossz helyre történő beillesztése 0,5, és túl nagy távolságra (legalább három szó) történő rossz helyre való beillesztése pedig 1 pont. Az elemzett szövegrészen végighaladva a pontokat az összegzés után a jelölők számával osztjuk, a felső korlátjaként pedig (a többi mérőszám felső értékét figyelembe véve) 1-et adtunk meg.

Végül, a kifejezés utolsó tagja $I_{Ti}(T_1, T_2)$ az időbélyegek pontosságát összehasonlító mérőszám, melynek értékét az időbélyegek átlagos, egymástól való eltérése adja. A munka jellegéből adódóan nem várjuk el, hogy az időbélyegek helye teljesen egybeessen, így a három másodperc alatti eltéréseket egyező időbélyegnek tekintjük. Ezt azért engedhettük meg, mert az egyes szövegrészeket az időbélyegek elválasztják, és adott esetben a mondatok befejezését követően a következő beszélő mondandójának megkezdése közé bárhová eshet időbélyeg. Az efeletti pontatlanságokhoz tartozó pontszámot az alábbi összefüggéssel számítjuk (alább láthatjuk a kifejezést szemléltető diagramot):

$$I_{Ti}(t_1.t_2) = 1 - e^{-0.25(|t_2 - t_1| - 3)} \quad (2)$$

[2] A Levenshtein távolság a szövegrész teljes hosszának az arányában.

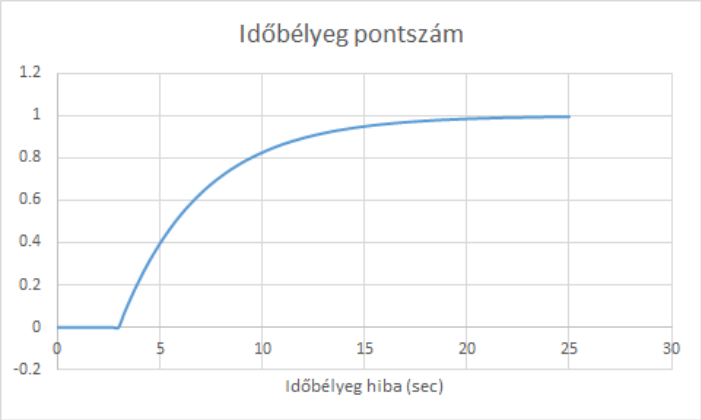

Fig. 2. Az időbélyegek eltéréséből számolt hibapontszám

Ez a kifejezés az egyedi időbélyeg értékek összehasonlítását adja meg. Azaz eszerint az időbélyegek esetében sem "nézzük el" az alacsony hibákat, hiszen a pontszám már 10 másodperces hibáért (egy egyórás szövegben ez egyszerűen bekövetkezhet, hiszen a szöveg előrehaladtával az időbélyegekben vételt hiba kumulálódik) is igen magas értéket vesz fel. Mivel azonban az egyszeri csúszás a gépelés során gyakorlatilag egy állandóan magas értéket eredményezne, hiszen a csúszást követően minden időbélyeg torzulhat, figyelembe kell vennünk az adott szegmensek hosszát is. A szegmenshosszok hibáit az időbélyeghez rendelt hibapontszámnál figyelembe véve tehát az alábbi összefüggést használtuk.

$$I_{Ti}(T_1.T_2) = 0.1*\left(1-e^{-0.25(|t_{k2}-t_{k1}|-3)}\right)+0.8*\left(1-e^{-0.175(|t_{h2}-t_{h1}|-3)}\right) + 0.1*\left(1-e^{-0.25(|t_{z2}-t_{z1}|-3)}\right) \quad (8)$$

Ez a súlyozott exponenciális kifejezésekből álló összeg akkor vesz fel 1-hez közeli értéket, hogyha az időbélyegekben csúszás van (3 mp-nél nagyobb, a kezdő illetve záró időbélyegeket a „t_{kx}", illetve „t_{zx}" jelöli), illetve ezzel egyúttal a szegmenshossz sem megfelelő (utóbbi szerepel a legnagyobb súllyal, a szegmenshosszokat „t_{hx}" jelöli). Ezáltal a gépelés során jelen lévő állandó csúszást jóval kisebb mértékben büntetjük, mint ha a feldolgozó egy adott szegmenst nem megfelelő hosszúságúnak rögzít. Ettől eltérő módszerekkel is meghatározható az időbélyegekre adott pontszám, azonban az itt leírt szempontok figyelembe vétele számunkra a minőségbiztosításban elegendő.

Végül az (1) kifejezés szerint az IRI különböző elemeit különböző súlyokkal vesszük figyelembe (melyek összege 1-et ad) attól függően, hogy melyik pontosságra helyezzük a legnagyobb hangsúlyt. Emellett célszerű alacsony súlyt adni az időbélyegek pontosságát leíró mennyiségnek, mivel ez a többivel ellentétben nem normált mennyiség. Ennek megfelelően, ha a többi mérőszámmal megegyező súllyal vesszük figyelembe, azzal torzíthatjuk az IRI index értékét. Jelen írásunkhoz a 0.2-0.6-0.2 súlyokat használtunk, kiemelve a jelző pontosságának számítását.

Az ismertetett megoldás egy viszonylag egyszerű lehetőséget ad a feldolgozók szövegeinek az összehasonlítására, azonban szem előtt kell tartanunk, hogy az egyszerű számként létrejövő IRI nem feltétlenül minden esetben elég informatív a munka javításához. Emellett számos probléma felmerülhet a szövegekkel kapcsolatban, amelyek az IRI-t kiegészítő megoldásokat kívánnak.

5.2.2. Praktikus megfontolások és az IRI alkalmazása

A gyakorlatban az IRI-t úgy használhatjuk, hogy egy adott szövegrészt több feldolgozóval is legépeltetünk és annotáltatunk, majd az így létrejövő szövegekre egyedileg kiszámoljuk az IRI értékét. Praktikus okokból nem csupán a kompozit index számítását végezzük el, hanem az IRI részeit külön-külön is elemezzük, hogy megértsük, hogy az egyes feldolgozók esetében mely feladatok jelentik a problémát. Ez főként a tagek alkalmazásánál, és az időbélyegek elhelyezésénél fontos.

A tagek esetében az IRI kiszámítását mindig adott jelölőkategóriákra végezzük el, amellyel aggregált értéket kapunk. Az egyes kategóriák egyedi elemzése lehetővé teszi közülük a problematikusak azonosítását, így például a különböző típusú hanghatások észlelhetőségét, a beszélgetésekben jelen lévő harmadik személy jelenlétét, stb. Továbbá külön vizsgálhatóak a hibakategóriák szerinti "pontszámok", így felismerhető, ha egy-egy feldolgozó valamely taget túl gyakran, vagy éppen túl ritkán használja, esetleg szisztematikusan rosszul helyez el a gépelt szövegben.

Az időbélyegek esetében elsőként a fentebbinél egyszerűbb mérőszámot vizsgálunk: az adott szövegben levő időbélyegek számát, és időbeli eltérését. Itt a feldolgozók számára értékes visszajelzés az, ha túl sok, vagy túl kevés időbélyeget látunk a munkájukban. Azonban az időbélyegek használatában való eltérés egyéb problémákra is rámutathat. A feldolgozók például gyakran eltérően és nem megfelelően szegmentálják velük a szöveget, vagy pedig olyan szövegrészeket is legépelnek, amelyet nem kellene (egy nem az adott diskurzusban résztvevő verbális és nem verbális közléseit). Nem ritka tehát, hogy két szövegben eltérő szegmenseket, illetve valamely szövegben indokolatlan szegmenseket találtunk. Az időbélyegek száma mellett a szegmenseket kezdő és záró időbélyegek pontossága is fontos mérőszám.

Emellett, mivel a minőségbiztosításra használt gépelt szöveg mérete megfelelt a beszélt szövegben az egy órás időtartamnak, az IRI számítási algoritmusának futásideje is igencsak megnőtt a nagyszámú természetes szegmens összerendelése miatt.

A feldolgozók munkájának összevetését tehát -- a fentebb elmondottak okán -- nem lehetett az időbélyegekre támaszkodva elvégezni, így a szöveg mesterséges szegmentálása mellett döntöttünk. A munkát az ún. horgonyszavakra támaszkodva végezzük. A horgonyszavak az összehasonlított szövegekben egyaránt maximálisan egyszer előforduló, négy betűnél hosszabb, értelmes szavak. Emellett egy olyan megkötést is tettünk, hogy csak azokban a természetes, (azaz a feldolgozók által létrehozott) szegmensekben keressük a horgonyszavakat, amelyeket az adott hangrögzítőt viselő személy mondott (azaz amelyek a beszélőhöz tartozó jelölővel kezdődtek)[3]. A mesterséges szegmentálást

[3] Ennek a megkötésnek a számítások felgyorsítása mellett a praktikus oka az volt, hogy a többi szereplő által elmondott szöveget a feldolgozók nagyon eltérő minőségben gépelték. Ebből fakadóan különböző hosszúságú szegmenseket kaptunk, és sok esetben egyes feldolgozók olyan szegmenseket is legépeltek, amit a többiek nem. Végül tehát a legépelt szövegekben a megegyező, egyértelműen összehasonlítható tartalom az adott beszélő szövege volt.

úgy végeztük, hogy az adott horgonyszavak feldolgozók által kijelölt mondatait tekintettük a szegmensek határainak – a feldolgozók jellemzően hasonló helyekre tették a mondatvégi írásjeleket a beszélt szövegben. Azaz egy adott mesterséges szegmens egy olyan mondattal kezdődött, amelyben benne volt az adott horgonyszó.

Így nagyobb számú, de kisméretű szövegszegmenst kaptunk. Vizsgálataink során azt tapasztaltuk, hogy az így létrehozott mesterséges szegmensek akkor is jól összerendezhetőek, ha az időbélyegekben, vagy akár a szegmensek hosszában komolyabb eltérés mutatkozik. Ez alapján az IRI indexeket a mesterséges szegmensek között számítjuk ki, a végső indexeket pedig a szegmenshosszal súlyozott összegként. Egy mesterséges szegmens nagyjából egy perces szöveget fogott át így.

5.3 Eredményeink az első minőségbiztosítási fázisban

Az első minőségbiztosítási fázis a korpuszépítés kezdeti szakaszában, közvetlenül a feldolgozási útmutató megismertetése és a feldolgozók betanítása után zajlott. A méréshez minden feldolgozónak ugyanazt a szöveget adtuk. A szöveg a teljes hanganyag kis része volt, amelyet úgy választottunk meg, hogy a feldolgozóknak az összes típusú jelölőt használniuk kelljen, illetve, amelyben kellő mennyiségű időbélyeg is szerepel.

Fontos lépés volt a referenciaszöveg kiválasztása is. Erről a rendelkezésre álló szövegek kézi feldolgozásával, a minőség manuális vizsgálata alapján döntöttünk, különös figyelmet fordítva a jelölők helyes alkalmazására.

A 2. táblázatban először is a szövegek struktúráját tekintjük át. Amint arra a táblázat adatai rámutatnak, a szövegek hossza feldolgozónként változik. Néhol ez az eltérés számottevő, ami annak tudható be, hogy a feldolgozók a távolabbi résztvevők mondatait eltérő mértékben érzékelik. Ez hasonlóképp igaz az időbélyegek számára, illetve az alkalmazott jelölőkre is. A táblázatban az időbélyegek és a jelölők száma mellett látható a legépelt szavak száma is. A referenciaszövegként használt szövegben található a legtöbb időbélyeg és jelölő, így a jelölők összehasonlításakor kevésbé fordulhat elő, hogy a referenciaszöveg a hiányos.

2. Táblázat: Az ellenőrzött szövegek gyakorisági jellemzői

Szöveg	Karakter-szám	Szavak száma	Szöveg	Karakter-szám
Referencia	64273	10202	1940	1357
Feldolgozó 1	64414	10276	1868	1139
Feldolgozó 2	33737	5183	911	637
Feldolgozó 3	57842	9001	1503	1105
Feldolgozó 4	43837	6482	1203	1051
Feldolgozó 5	54992	8951	1312	1132
Feldolgozó 6	48943	7326	1404	1079
Feldolgozó 7	57674	9323	1483	1306
Feldolgozó 8	60449	9760	1567	1413
Feldolgozó 9	52269	8187	1399	897

A szövegek horgonyszavas szegmentálását követően 55 szegmenst azonosítottunk, melyek átlagosan 380 karakterből állnak, azonban igen nagy szórásúak voltak (321 karakter).

Az elemzés következő lépéseként kiszámítottuk az IRI pontszámokat az egyes szövegekre a referenciaszöveghez hasonlítva. Praktikus okokból a pontszámokat kategóriánként számítottuk a végleges IRI pontszám ebből az (1) kifejezésben említett súlyozás szerint adódhat. A minőségbiztosítási folyamatban ezeket a pontszámokat külön-külön alkalmaztuk. Az eredmények azt mutatják, hogy a feldolgozók egyaránt konzisztens teljesítmény nyújtottak mind a szöveges tartalmak egyezését illetően (tehát, amelyben nem voltak benne a jelölők és az időbélyegek), mind az időbélyegek és a jelölők alkalmazásának terén. Ezért végül csupán a számunkra az annotálás szempontjából legfontosabb jellemzőt vettük alapul a feldolgozók rangsorolásában: a jelölők használatát[4].

3. Táblázat: A szövegekre számolt IRI részértékek

Szöveg	Szóbeli egyezés	Időbélyeg pontszám	Jelölő pontszám
Feldolgozó 1	0.643885701	0.425844137	0.513876
Feldolgozó 2	0.671654333	0.264923636	0.456804
Feldolgozó 3	0.641393827	0.257121049	0.531356
Feldolgozó 4	0.681724923	0.243786923	0.497431
Feldolgozó 5	0.699152208	0.212764862	0.585468
Feldolgozó 6	0.669104015	0.228939292	0.509069
Feldolgozó 7	0.640636905	0.237307279	0.519228
Feldolgozó 8	0.713745273	0.218149973	0.531762
Feldolgozó 9	0.720822972	0.167920389	0.54657

Az egyszerű rangsorolást követően vettük figyelembe az időbélyegek és a szóbeli egyezés mutatóit, majd részletekbe menően vizsgáltuk a jelölők használatát jelölőkategóriák szerint is, itt azonban nem célunk ennek ismertetése.

A minőségbiztosítási fázis végén a teljesítmény alapján hat feldolgozóval nem folytattuk tovább a munkát, a többiekkel pedig részleteiben egyeztettük a minőség-ellenőrzés eredményét. Ennek eredményeképp sikerült javítanunk a tagek használati szabályait, tehát sokkal precízebb kritériumokat megadni az annotálás részleteivel kapcsolatban.

6. Összegzés, további tervezett lépések

Dolgozatunkban a jelenleg is fejlesztés alatt lévő HuTongue, magyar nyelvű spontán beszéd korpusz minőségbiztosításában alkalmazott néhány módszert mutattuk be. A korpusz méretéből fakadóan a gépelés és az annotálás komoly humán munkaerőt igényel, így a megfelelően alapos és pontos minőségbiztosítás elengedhetetlen. Erre az

[4] Ez alól egyedüli kivétel az 1. Feldolgozó volt, ahol a többiektől elmaradó szövegi egyezésnek és a magas időbélyeg pontszámnak köszönhetően egyedileg is elemeztük a hibákat.

általánosan használt módszerek mellett egy, a HuTongue különleges jellemzőit figyelembe vevő egyedi mérőszámot alkottunk, amellyel jelentősen egyszerűbbé tettük a minőségbiztosítási fázisokban gépelt dokumentumok elemzését és a gépelést végzők teljesítményének az értékelését.

A dolgozatban röviden ismertettük az általunk definiált mérőszám részleteit, valamint az alkalmazásánál figyelembe vett gyakorlati szempontokat. Ezt követően, az első minőségbiztosítási fázis adatain bemutattuk, hogyan lehet a mutatót a gyakorlatban alkalmazni.

Az ismertetett eredmények alapján azt látjuk, hogy a mutató már jelenlegi formájában is alkalmas az egyes minőségbiztosítási fázisokban a feldolgozók értékelésére, azonban lehetséges, hogy további finomítással a mutató még érzékenyebbé tehető a feldolgozás során elkövetett hibákra. A jelenleg alkalmazott, a tagek és időbélyegek közti eltéréseket mutató mérőszámok korlátozott mértékben érzékenyek bizonyos hibákra. Az időbélyegek esetén a hibák típusának átsúlyozása érzékenyebben mutathatja a hibákat, illetve a jelölők használata esetén lehetséges, hogy nyers mutatók bevétele a mutatók számításába (például az egyes szegmensekben a beillesztett jelölők közti számszerű különbség) is javíthatja az IRI-t. Emellett a jelenleg használt Levenshtein távolságot egy jóval részletesebb mérőszámmal is lehetne helyettesíteni, amely kevésbé érzékeny a szavak sorrendjére, és inkább az adott szegmensekben szereplő szavak és azok jelentése közti eltéréseket veszi figyelembe.

Mindezt összefoglalva azt tapasztaltuk, hogy a fejlesztett megoldás fontos előnye egyrészt az, hogy folyamatosan nyomon tudjuk követni a munka minőségének szinvonalát, másrészt a kapott mérési eredmények alapján visszajelzésre vagyunk képesek a feldolgozók felé, biztosítva ezzel a munka minőségének folyamatos javítását. Ennek köszönhetően amellett, hogy a gépelés pontossága javult, az annotációs jelölők használata is jelentős mértékben fejlődött.

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében (ERC_CoG_2014_648693 sz. szerződésben) támogatja, a kutatás vezetője Takács Károly.

Bibliográfia

1. Crowdy, S.: Spoken Corpus Design. Lit Linguist Computing (1993) **8**(4): 259–265.
2. Galántai J., Pápay B., Kubik B., Szabó M., és Takács K.: A pletyka a társas rend szolgálatában: Az informális kommunikáció struktúrájának mélyebb megértéséért a Computational Social Science eszközeivel. Magyar Tudomány. Megjelenés előtt.
3. Gósy M., Gráczi T.E., Gyarmathy D., Váradi, T., Veresné Horváth, V.: Magyar spontán beszéd adatbázis = Hungarian Spontaneous Speech Corpus. Munkabeszámoló. OTKA (2012) (http://www.nytud.hu/adatb/bea/index.html).

4. Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In Proceeding HLT '90 Proceedings of the workshop on Speech and Natural Language. Hidden Valley, Pennsylvania. (1990) 96–101.
5. Kugler N.: Megfigyelés és következtetés a nyelvi tevékenységben. Budapest, Tinta. (2015)
6. Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In Proceedings of LREC. (2000) 947–952.
7. Oostdijk, N.: The spoken Dutch corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer eds. Proceedings of the Second International Conference on Language Resources and Evaluation 2. (2000) Paris. ELRA. 887–893.
8. Szabó, M.K., Galántai J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In MANYE-kongresszus konfenreciakötete (2017) Megjelenés előtt.