Nagy C. Katalin:
Adatforrások integrációja a történeti nyelvészetben: a nyelvtörténeti korpuszadat fogalmáról
Argumentum 9 (2013), 56-78
Debreceni Egyetemi Kiadó

Tanulmány

Nagy C. Katalin

Adatforrások integrációja a történeti nyelvészetben: a nyelvtörténeti korpuszadat fogalmáról*

Abstract

The present paper aims to clarify the notion of 'historical corpus data' from a methodological point of view, through the analysis of examples taken from actual research practice. Relying on Kertész and Rákosi's (2012) p-model of scientific theorizing, it presents a novel conception of historical data, which rejects the use of the term 'historical data' as a synonym for occurrences found in a historical corpus. The paper concludes that historical data should be conceived of as plausible statements originating from an integrated data source. Historical corpus data, as they appear in historical linguistic research and argumentation, integrate more information than the amount of information retrievable from the corpus alone. The analysis of historical research practice suggests that the source of historical corpus data is of an integrated nature. The present paper offers a notion of historical corpus data suitable to show the uncertainties concerning the use of historical corpora and the integrated nature of the source where data used in historical research originate from.
Keywords: historical data, corpus data, data sources, integration of data sources

1 Bevezetés

A nyelvészeti adatok kérdése a közelmúltban a metanyelvészeti gondolkodás előterébe került (vö. Lehmann 2004; Kepser & Reis 2005; Penke & Rosenbach 2004/2007; Kertész & Rákosi 2008a, 2008b, 2012). Az adatkérdéssel kapcsolatos vizsgálódások egyik konklúziója azon álláspont túlhaladottsága, hogy a nyelvészeti kutatás során elegendő lenne kizárólag egy adattípusra alapozni, legyen az korpuszadat vagy introspekcióból származó adat (l. Wasow & Arnold 2005; Kertész & Rákosi 2008c, 2008d). A különböző adattípusok integrációjának elméleti kérdései azonban még nem megfelelően tisztázottak (l. Kertész & Rákosi 2013).

A különböző adattípusok kombinálása a történeti nyelvészetben is kívánatos, bár ezen a kutatási területen mindig is megfigyelhető volt a korpuszadatok – részben szükség szülte – központi szerepe. Azoknak a nyelvállapotoknak a vizsgálatában, amelyekből már fennmaradtak írásos dokumentumok, a többi adatforrás csupán a történeti korpusz kiegészítőjeként jelenik meg. Ennek ellenére a 'történeti korpuszadat' fogalma a nyelvészeti szakirodalomban nem megfelelően tisztázott: a 'történeti adat' kifejezés a történeti korpuszban lelt előfordulások

* Jelen tanulmány a MTA-DE Elméleti Nyelvészeti Kutatócsoport támogatásával jött létre. Ezúton fejezem ki köszönetemet Németh T. Enikőnek a tanulmányban szereplő gondolatok kibontásában nyújtott segítségéért, valamint Rákosi Csillának a tanulmány korábbi verziójához fűzött építő meglátásaiért.

Nagy C. Katalin:
Adatforrások integrációja a történeti nyelvészetben: a nyelvtörténeti korpuszadat fogalmáról
Argumentum 9 (2013), 56-78
Debreceni Egyetemi Kiadó

szinonimájaként jelenik meg. A történeti korpuszban lelt előfordulások felhasználása azonban csak akkor lehetséges, ha számos egyéb adatforrást is bevonunk a vizsgálatba. Így az adattípusok integrációja a kutatás megfigyelhető gyakorlataként és szükségszerűségeként a történeti kutatásban is jelen van, a történeti korpuszon alapuló kutatások ugyanis semmiképpen sem tekinthetőek egyetlen adatforrást használó kutatásnak.

Írásomban a korpuszhasználat egyes kérdéseit vizsgálom meg a történeti nyelvészetben, a következő problémakörök köré rendezve. (i) Hogyan értelmezhető a 'történeti korpuszadat fogalma?[1] (ii) Mik a történeti korpuszadat felhasználásának korlátai? (iii) A különböző adattípusok integrációja hogyan jelenik meg a korpuszadatok felhasználásakor? Tanulmányom a következőképpen épül fel. A történeti kutatás adatforrásainak áttekintése után (2. pont), a 3. pontban megvizsgálom a történeti korpuszadat fogalmának hagyományos értelmezését és az azzal kapcsolatban felmerülő problémákat. Ezeknek a problémáknak a kiküszöbölésére a történeti korpuszadat fogalmát a Kertész és Rákosi által (2012) kidolgozott p-modell adatfogalma alapján értelmezem újra. A 4. pontban sorra veszem a történeti korpusz felhasználásakor felmerülő módszertani problémákat. A korpuszhasználatot illető módszertani megfontolások ahhoz a megállapításhoz vezetnek, hogy a történeti adat nem egyetlen forrásból, hanem források összességéből, azaz integrált adatforrásból származó adatnak tekinthető. Ezt az elképzelést az 5. pontban mutatom be, majd megvizsgálom ennek az integrációnak a természetét (6. és 7. pont). Az utolsó, 8. pont a tanulmány összefoglalását tartalmazza.

2 A történeti kutatás adatforrásai

Bár Jucker (2009: 1615–1619) a pragmatika módszereiről írva különítette el a nyelvi intuíción alapuló adatok, a természetes nyelvhasználat adatai és előidézett/kísérleti adatok[2] három nagy csoportját, ezeket a nyelvészeti kutatás más területeit illetően is tekinthetjük az adatok három alaptípusának. A nyelvi intuíción alapuló adatok forrása lehet magának a kutatónak a nyelvi intuíciója, de az interjú módszerrel más anyanyelvi beszélők nyelvi intuícióján alapuló adatok is gyűjthetőek, ide tartoznak az egyes nyelvi szerkezetek grammatikalitására rákérdező kísérletek is. A természetes nyelvhasználat adatai a jegyzetfüzet módszerrel begyűjtött, mindennapi életben előforduló nyelvi példákon alapuló adatok, a filológiai módszerrel szerzett, irodalmi vagy egyéb írásművekből származó előfordulások, a konverzációelemzés módszerével begyűjtött tényleges társalgások átírásai, valamint a korpuszmódszerrel gyűjtött releváns előfordulások. Végül, az előidézett vagy kísérleti adatok többek között szóban vagy írásban végrehajtott diskurzus kiegészítéses és feleletválasztós tesztekkel vagy szerepjátékokkal gyűjthetőek. Ez utóbbi módszernél is természetesen szerepet játszik az adatközlők nyelvi intuíciója, hiszen a kutató arról kér információt, hogyan viselkednének egy-egy elképzelt kommunikációs helyzetben. Az adatközlők véleménye azonban, szemben az első csoporttal, itt nem közvetlenül nyilvánul meg: az adatok az elképzelt kommunikációs szituációban használt megnyilatkozásokon alapulnak. Bár az így előidézett kommunikációs helyzetek mesterségesnek hathatnak, ez a módszer a kutató számára nagyobb kontrollt tesz lehetővé a különböző változók fölött.

---

[1] A nyelvtörténeti adat problémakörének, ezen belül pedig a magyar nyelvtörténeti vonatkozásoknak egyes kérdéseit tárgyalja Dömötör (2012).

[2] Az *adat* terminus hagyományos értelemben vett használata különböző elméleti és módszertani kérdéseket vet fel, amelyekre később térek majd ki. Addig az adat intuitív fogalmát használom, ahogyan az a hivatkozott szakirodalomban megjelenik.

A fent említett adattípusok közül nem mindegyik érhető el a történeti kutatás számára. A kutató nyelvi intuíciója mint adatforrás, ahogyan mindenféle nyelvészeti kutatásban, úgy a történeti kutatásban is szükségszerűen jelen van. A régebbi korok nyelvére vonatkozó nyelvi ítéletek meghozatalához azonban ún. pótkompetencia kifejlesztése szükséges (l. később a 4.4. pontban). Más anyanyelvi beszélők nyelvi intuíciójáról a történeti nyelvész csupán áttételesen juthat információhoz: ilyennek számítanak a korabeli nyelvtanírók, nyelvművelők munkáiban, esetleg nyelvkönyvekben fennmaradt normatív jellegű megjegyzések, vagy a kéziratok későbbi másolataiban fellelhető eltérések az eredetitől, margóra írt megjegyzések, fordítások stb. A rendelkezésre álló metalingvisztikai információk mennyisége azonban az egyes történeti kutatásokban változó lehet, sőt teljesen hiányozhat. A jegyzetfüzet módszer és a konverzációelemzés ugyan korábbi nyelvállapotok vizsgálatában nem használható, de a filológiai módszer és a korpuszmódszer elérhető azon nyelvállapotok tanulmányozói számára, amelyekből már fennmaradtak írott források. Az olyan nyelvállapotokra vonatkozó nyelvtörténeti kutatások, amelyekre nézve nincsenek írott forrásaink, csupán rekonstruált adatokra építhetnek. Előidézett adatok gyűjtésére a történeti kutatásban anyanyelvi beszélők hiányában nincs lehetőség.

Végül érdemes megjegyezni, hogy a szinkrón kutatás azon adatforrásai, amelyek a történeti kutatásban nem elérhetőek, közvetett módon mégis szerepet játszhatnak a diakrón kutatásban, hiszen a történeti nyelvészeti érvelés is használhat a szinkrón nyelv vizsgálatán alapuló ismereteket.

Amint a fenti megfontolásokból kitűnik, a történeti kutatás adatforrásai a szinkrón kutatásénál korlátozottabbak, ezért a történeti nyelvészeti hipotézisek empirikus alátámasztottsága is problematikusabb. Ezt a problémát a szövegstílusok összevetésével (formális-informális), a beszélt nyelven alapuló írott források (pl. politikai beszédek, prédikációk, személyes levelezés) bevonásával, valamint metalingvisztikai adatok (illemtankönyvek, nyelvkönyvek, nyelvészeti munkák, egyéb, társalgásról szóló leírások) felhasználásával lehet némileg ellensúlyozni (l. Jacobs & Jucker 1995; Fischer 2007). Felhasználhatunk ezen kívül tipológiai adatokat, a nyelvcsaládra vonatkozó, belső vagy összehasonlító rekonstrukción alapuló információkat (Fischer 2007: 44), a szinkrón nyelvből nyert, nyelvi változatossággal kapcsolatos információkat és magából a grammatikalizációs elméletből eredő adatokat is (ehhez l. Heine 2005 [2003]: 580, 585–586). Fischer (2004, 2007: 15–17) azt javasolja a történeti nyelvészet adatforrásainak bővítésére, hogy legyünk érzékenyek a teljes nyelvállapotra: ne csak azt az elemet vizsgáljuk, amely a kutatás kimondott tárgya, hanem más, funkcionálisan és formailag hasonló elemeket is. Végül, Fischer szerint (2007: 44) az abból adódó hiányt, hogy az anyanyelvi beszélők kompetenciája már nem elérhető, a kvantitatív információkra való erőteljesebb támaszkodással ellensúlyozhatjuk. Fischer javaslata kétségkívül hasznos eljárás, figyelembe kell azonban vennünk, hogy a gyakori előfordulás nem lehet követelmény egy adott szerkezet grammatikusságának megállapításakor.

Összefoglalóan, a történeti kutatásban a következő adatforrásokkal számolhatunk: történeti korpusz, a vizsgált nyelvállapothoz képest korábbi (akár rekonstruált) vagy későbbi (akár a jelenlegi) nyelvállapotra vonatkozó információk, metalingvisztikai források, tipológiai források, maga a felhasznált elmélet, a kutató nyelvész intuíciója, enciklopédikus ismeretek. Mindezek között központi jelentőséggel bír a történeti korpuszon alapuló adat. Azokra a nyelvállapotokra nézve, amelyekből fennmaradtak írott dokumentumok, a többi adatforrás a történeti korpuszadatot kiegészítve jelenik meg a kutatásban. Tanulmányomban tehát erre az adattípusra koncentrálok, és a korpuszadatok mibenlétét, valamint a korpuszt mint adatforrást vizsgálom meg.

Nagy C. Katalin:
Adatforrások integrációja a történeti nyelvészetben: a nyelvtörténeti korpuszadat fogalmáról
Argumentum 9 (2013), 56-78
Debreceni Egyetemi Kiadó

3 A korpuszadat fogalmáról

3.1 A történeti korpuszadat hagyományos megközelítése

A történeti dokumentumokból gyűjtött előfordulások központi szerepe a történeti nyelvészeti kutatásban kétségbevonhatatlan. Fischer (2004: 730) a történeti dokumentumokat egyenesen „a történeti nyelvész ismereteinek egyetlen biztos forrásaként”[3] nevezi meg. Lehmann (2004: 201) – bár az ő megállapítása nemcsak a történeti kutatásra vonatkozik –, szintén a korpuszon alapuló adatokat tekinti a legmegbízhatóbb adattípusnak. A szerző szerint a korpuszadatok „talált” jellegüknél fogva objektívebb kutatást tesznek lehetővé, hiszen függetlenek a kutatótól, így kevesebb mód nyílik az adatok manipulálására. A talált korpusz változatos, meglepetésekkel szolgálhat, új, meglepő, előre nem látott felfedezésekre ad módot. Fischer és Lehmann optimista hozzáállását tükrözi a történeti kutatás gyakorlata is: a kellően kiterjedt korpuszt felhasználó kutatások nagy tekintélynek örvendenek. A korpuszbeli előfordulások fontosságának elismerése mellett megkérdőjelezhető, hogy a korpusz valóban ismeretek biztos, objektív eredményekhez vezető forrásának tekinthető-e (vö. Kertész & Rákosi 2012: 173–175). Mielőtt azonban rátérnénk erre a kérdésre, azt kell megvizsgálnunk, mi is az a történeti korpuszadat.

A történeti kutatásokban forrásként valamely történeti korpuszt szokás megjelölni. Francis (1992: 17) a nyelvészeti korpuszt így definiálja: „szövegek egy adott nyelvre, vagy annak egy részére nézve reprezentatívnak[4] tartott, nyelvi elemzésre használható gyűjteménye”[5]. A történeti korpuszt képező szövegek halmaza jelöli ki a kutatás adatait, ami alatt a hagyományosan a vizsgált jelenséget tartalmazó előfordulásokat értik. Ennek megfelelően a történeti kutatások bemutatásakor a ’történeti adat’ kifejezés az előfordulásokra vonatkoztatva jelenik meg. Ezt a felfogást mutatja a következő, általános gyakorlatot tükröző szóhasználat, amellyel Juge (2006) a katalán „*anar* + főnévi igenév” szerkezet grammatikalizációja morfológiai vonatkozásainak feltárását célzó kutatását vezeti be: „a kérdést kettő, a legkorábbi és legjelentősebb katalán narratív szövegek közé tartozó szövegből, I. Jakab *Llibre dels Fets* című művéből és Bernat Desclot *Krónikájából* vett adatokon vizsgálom meg”[6] (Juge 2006: 318). Ebben a megfogalmazásban az adat kifejezésen feltehetőleg a kérdéses előfordulások értendőek.

Ha azonban az *adat* terminust az *előfordulás* szinonimájaként értjük, akkor nem tudjuk leírni, hogyan is épülhet a kutatás adatokra, amire ugyanis a kutatás és érvelés során építünk, azok valójában az előfordulásokról tett *kijelentések* (Kertész & Rákosi 2012: 170–171). Ráadásul ahhoz, hogy egy előfordulásról tett kijelentést be tudjunk kapcsolni a nyelvészeti elméletalkotás folyamatába, további adatok bevonása szükséges. Így például (Kertész & Rákosi 2012: 169–170) kifejti, hogy amennyiben adaton csupán annyit értünk, hogy egy mondat vagy szósor egy adott nyelv része, vagy megtalálható egy korpuszban, akkor egy további adatot is be kell vonnunk, amely e nyelvi kifejezés szerkezetéről állít valamit. Vajon hogyan válik a korpuszban talált előfordulásokból a kutatásban felhasználható adat? A továbbiakban az adat

---

[3] „The historical linguist has only one firm source of knowledge and that is the historical documents.”

[4] A korpusz reprezentativitásának kérdése természetesen a történeti korpuszokkal kapcsolatosan is felmerül, ezzel azonban jelen tanulmányban nem foglalkozom.

[5] „a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis”

[6] „I examine the issue with data from the *Llibre dels Fets (Book of Deeds)* of King Jaume I (1313–1327) and Bernat Desclot’s *Crònica (Chronicle)* (1283–1295, chs. 1–106), two of the earliest and most important Catalan narrative texts.”

fogalmának egy olyan megközelítését mutatom be, amelynek segítségével a történeti korpuszadatoknak a kutatás folyamatában játszott szerepe jobban leírható.

3.2 Az adat fogalma a p-modellben

A Kertész és Rákosi (2012) által kidolgozott ún. p-modell hasznos kerete lehet a történeti adat olyan felfogásának kialakításához, miszerint az egy történeti dokumentumban lelt előfordulásról tett kijelentés. Kertész és Rákosi p-modelljének elnevezése a modell központi gondolatára utal: a szerzők szerint a tudományos elméletalkotás plauzibilis érvelési folyamat: csupán részlegesen alátámasztott, plauzibilis információkra épül és plauzibilis eredményekhez vezet (Kertész & Rákosi 2012: 54). A szerzők ennek megfelelően az *adat* kifejezést nem a hagyományos, fentebb bemutatott értelemben használják. Újszerű adatfogalmuk tükrözi, hogy a kutatás során felhasznált információk nem tekinthetőek teljesen bizonyos ismereteknek, csupán plauzibilisnek. A p-modell ugyanis adatokon plauzibilis kijelentéseket ért, amelyek szerkezete két komponensből áll: az információtartalmukon túl egy plauzibilitási értéket is tartalmaznak, amely valamely forrás alapján rendelhető hozzájuk. Az adatok kiinduló plauzibilitási értéküket valamely direkt adatforrásból kapják, Kertész és Rákosi (2012: 64) Rescher (1976: 6) alapján többek között a következőket említi: érzékszervi tapasztalatok, szakértők, szemtanúk, történelmi források, az emlékezetünk, sejtések, feltevések, korpuszok, elméletek, bizonyos elvek (pl. egyszerűség, egyöntetűség stb.), egyaránt szolgálhatnak direkt forrásként.

A kijelentések plauzibilitási értékét kezdetben forrásuk megbízhatósága határozza meg, azonban az az érvelés folyamán megváltozhat, amennyiben forrásuk megbízhatósága megkérdőjeleződik. A p-modell azt is megengedi, hogy egy kijelentés egyidejűleg legyen plauzibilis valamely forrás alapján, míg egy másik alapján implauzibilis.

A fenti megfontolások alapján az adatok Jucker (2009: 1611, 1615–1619) által elkülönített három csoportja – a nyelvhasználók nyelvi intuícióján alapuló, a természetes nyelvhasználatból vagy filológiai módszerrel gyűjtött adatok és az előidézett vagy kísérleti adatok – egységesen kezelhető. Ezek mind a kutató által vizsgált előfordulásokra vonatkozó kijelentésekként értendőek, amelyek információtartalmukon kívül plauzibilitási értékkel is rendelkeznek. Kiinduló plauzibilitási értéküket valamely direkt forrás alapján kapják (nyelvi intuíció, beszélt vagy írott korpusz, kísérlet).

3.3 A nyelvtörténeti korpuszadat fogalma a p-modellben

A fentiek alapján tekintsük, hogyan képzelhetőek el a korpuszadatok, köztük a történeti korpuszadatok a p-modell adatfogalmának megfelelően! Kertész és Rákosi modelljével összhangban a korpuszon alapuló adatok olyan plauzibilitási értékkel rendelkező kijelentések, amelyek egy kifejezés jelenlétéről vagy valamely tulajdonságáról tesznek állítást egy adott korpuszban. Ahogyan Kertész és Rákosi (2012: 173–175) megmutatja, szerkezetük különböző lehet: pl. „az X nyelvi jelenséget tartalmazó M megnyilatkozás megtalálható K korpuszban”, „az X szerkezet beazonosítható K korpuszban és rendelkezik Y tulajdonsággal”. Mivel a korpuszadatok is direkt forrásból származó plauzibilis kijelentésekként értelmezhetőek, nem tekinthetjük őket bizonyosan igaznak. A korpuszadatok megbízhatóságát számos tényező befolyásolja, többek között a dokumentum megbízhatósága, a kutató nyelvi intuíciója, a vizsgált jelenség/kifejezés interpretációjának nehézségei stb. (Kertész & Rákosi 2012: 173; vö. Forgács 1993–1994). Így a plauzibilis kijelentésekként felfogott korpuszadatok plauzibilitásának forrá-

sa nemcsak maga a korpusz, hanem több forrás együttese: maga a történeti szöveg, a nyelvész nyelvi intuíciója, a jártasság az adott elmélet alkalmazásában (l. később az 5. pontban). Mindezeknek a megbízhatósága befolyásolja az adat plauzibilitását.

4 A korpusz felhasználásával kapcsolatos problémák

A korpusz nem tekinthető ismeretek biztos, objektív eredményekhez vezető forrásának, felhasználása számos módszertani problémát vet fel, amelyeknek vizsgálatához a Kertész és Rákosi által (2008b, 2012) kidolgozott ún. *p-modellt* veszem alapul (a p-modell kapcsolódó kérdéseinek bemutatáshoz l. a kötet bevezető tanulmányát). A korpusz mint adatforrás használatával kapcsolatosan Kertész és Rákosi (2012: 173–175) négy problémát vet fel, amelyeket saját meglátásaimmal kiegészítve és a történeti korpuszadatokra adaptálva tárgyalok. A négy probléma a következő. 1. A korpusz felhasználásakor a kutatónak el kell tudni különíteni a grammatikus és a nem grammatikus formákat. 2. Az adatok között szelektálnia kell, ami elméletfüggőséghez vezet. 3. A korpusz felhasználásakor szükségképpen használnia kell nyelvi intuícióját. 4. Végül, a korpuszokkal kapcsolatosan felmerül a pozitív és negatív evidencia kérdése. E négy probléma ahhoz vezet, hogy a tudományos érvelés folyamatába bekerülő, korpuszon alapuló ismeretek nem tekinthetőek bizonyosan igaz információknak. A nyelvtörténeti korpuszadat hagyományos fogalma azonban nem tükrözi ezeket a bizonytalanságokat. A továbbiakban ezért az *adat* kifejezést a p-modell adatfogalmának megfelelően értem, amelynek segítségével értelmezhetőekké válnak a korpuszon alapuló adatokkal kapcsolatban felmerülő problémák. A továbbiakban tekintsük a fenti négy problémakört részletesebben!

4.1 A grammatikus és a nem grammatikus elkülönítése

A korpusz használata során, amint Kertész és Rákosi (2012: 174) felhívják rá a figyelmet, a kutatónak először is el kell döntenie, hogy az, amit a korpuszban talált, része-e a vizsgált nyelvnek, azaz el kell tudnia különíteni a grammatikus előfordulásokat pl. az elírásoktól. A történeti nyelvészetben ez a probléma összetettebben jelentkezik. Egyrészt, míg a mai nyelvállapot kutatásában lehetőség van a korpuszadatok korlátainak bizonyos fokú kiküszöbölésére, és az anyanyelvi beszélőkkel való konzultáció révén annak tesztelésére, hogy egy-egy, a korpuszban lelt „gyanús” előfordulás része-e a vizsgált nyelvnek, a történeti nyelvészetről ez nem mondható el. Másrészt, a történeti korpusz felhasználásakor nem elég eldönteni, hogy egy ott talált előfordulás része-e a vizsgált nyelvnek, hanem azt is mérlegelni kell, hogy része-e a vizsgált nyelvállapotnak. Egyazon dokumentumban ugyanis előfordulhatnak későbbi vagy korábbi nyelvállapothoz tartozó formák is, így a történeti szöveg nem feltétlenül (csak) azt a nyelvállapotot tükrözi, amikor íródott. Egy későbbi vagy korábbi nyelvállapothoz tartozó nyelvi elemre vonatkozó megállapítás természetesen adatként szolgálhat egy arra a másik nyelvállapotra vonatkozó kutatásban. A történeti szöveg az elírások mellett tartalmazhat archaikus formákat, más nyelvből vett szavakat, szövegrészeket, valamint később betoldott alakokat, javításokat is.

További nehézséget jelent, hogy az írott sztenderd megszilárdulása előtt lejegyzett történeti dokumentumokban az egyes nyelvi elemek írásmódja igen nagy változatosságot mutathat. Az írásmódok nagy változatossága, sokszor egyéni jellege, megnehezíti az esetleges hibák, elírások tettenérését, hiszen azokat el kell tudni különíteni a szándékosan másképp írott formáktól,

másképp fogalmazva, fel kell ismerni, hogy a történeti dokumentumokban gyakran tapasztalt különböző írásmódok konkrét megnyilvánulása elírás-e vagy pedig szándékosan-másképp-írás. Bár a különböző írásmódok megléte sok esetben egyszerűen az írott sztenderd hiányának tudható be, közülük egyesek utalhatnak valamilyen hangváltozásra is, ami egy történeti fonológus számára értékes információkat szolgáltathat.

Az elírás gyanúja a kutató nyelvi intuíciója alapján merül fel, így először is felveti a kérdést, hogy ennek megítélése milyen mértékben tekinthető megbízhatónak. Az elírásnak vannak viszonylag egyértelműnek tekinthető esetei, míg más esetekben annak eldöntése, hogy elírásról van-e szó, nem annyira egyszerű. Amennyiben valóban elírásról van szó, felmerül a kérdés, hogy vajon az eredeti kéziratban vagy a digitalizálás során keletkezett-e a hiba.

A történeti nyelvészetben sokkal gyakrabban kerülünk a fent leírt döntéshelyzet elé, bár a történeti kutatás egyes területeit jobban érinti ez a probléma, mint másokat. Mivel a hiba nagyobb valószínűséggel érint egyes szimbólumokat, mint egész szimbólumsorokat, a történeti fonológia számára kulcsfontosságúak a fenti problémákra adott válaszok, míg egy történeti pragmatikai vizsgálatban kevésbé súlyos tévedéshez vezet, ha például rosszul ítéljük meg, hogyan kellett kiejteni a vizsgált korban egy adott szót.

Az elírások tettenérésére vonatkozóan tekintsük a következő példákat.

(1) *tomo el cauallo por la rienda: & començo a* ***llamr*** *al infante diego gonçales. E el infante quando se oyo llamar por su nombre: torno la cabeça por ver quien lo llamaua.* (CrPC[7] 87v, középkori spanyol)
’megfogta a ló kantárát, és elkezdte **szólongatni** Diego Gonçales infánst. Az infáns pedig, amikor hallotta, hogy a nevén szólítják, megfordította a fejét, hogy lássa, ki szólongatja.’

A fenti, egy középkori spanyol krónikából származó szövegrészletben a *llamr* alak feltételezhetően elírásnak tekinthető szimbólumsor, ahol a helyes írásmód *llamar* lett volna. A második mássalhangzó hiányával létrejött forma által lejegyzett szóvégi mássalhangzó-kombináció a középkori spanyolban sem lett volna jól formált, és az elírás vélelmét az is megerősíti, hogy a kérdéses ige még kétszer előfordul a hibás forma közelében, ahol is a második mássalhangzó is megjelenik. Az alábbi eset azonban már nem ilyen egyértelmű. A ’hall’ igének a középkori spanyol szövegekben többféle írásmóddal dokumentált formái találhatóak meg, köztük az alábbiak:

(2) oir/**h**oir/oyr (< AUDIRE) ’hall’ (középkori spanyol)

(2)-ben látható, hogy a második írásmód esetében a szó elején egy *h* szimbólum jelenik meg. Amennyiben ezt egy szó eleji mássalhangzó szimbólumának tekintjük, az etimológiailag nem lesz levezethető, a kérdéses igealak ugyanis a latin AUDIRE igéből ered. Tudjuk azonban, hogy az *F- > h-* hangváltozás során a latin eredetű szókezdő *F-* hehezetté alakult, majd eltűnt a kiejtésből a spanyol nyelvben. A szó eleji *h* szimbólum ugyanakkor írásban sok esetben fennmaradt és még a mai spanyolban is használatos, de hangalakkal már nem rendelkezik. Amikor a vizsgált nyelvállapotban elkezd olyan szavak elején is megjelenni a *h* szimbólum,

[7] Crónica popular del Cid. [Cid népi krónikája]. In: *ADMYTE* (Archivo Digital de Manuscritos y Textos Españoles), CNUM 6993. 16. század eleje.

ahol eredetileg nem volt *f* hang, arra lehet következtetni, hogy a hangváltozás már lezajlott. A nyelvhasználók ugyanis már nem tudják megállapítani, hogy mely magánhangzóval kezdődő szavak elején volt eredetileg *f* hang, így tévesen kiteszik olyan szavak elejére is, ahol etimológiailag nem magyarázható, azaz hiperkorrekt írásmódokat hoznak létre. Ilyen értelemben a szókezdő *h* szimbólum alkalmazása a fenti példában nem elírásnak, hanem a szöveg lejegyzője szándékos tettének tekinthető, amely egy hangváltozás feltárásában és datálásában a kutató segítségére lehet. Amennyiben a szó eleji *h* szimbólumot egyszerűen elírásnak tekintjük, a történeti fonológia számára hasznos információkat veszíthetünk.

Az egységes írott norma kialakulása előtt lejegyzett szövegek esetén sok esetben nehézséget okozhat annak eldöntése, hogy a különböző írásmódok valamiféle hangváltozásra utalnak-e, vagy pedig egyszerűen a szó lejegyzőjének sajátos írásmódját tükrözik. Ennek megállapításában többek között az alábbi tényezők segíthetnek: a kutató nyelvi intuíciója, több előfordulás figyelembevétele, a „hiba" helye, a kutatónak az adott nyelvállapot fonetikai/fonológiai rendszeréről való ismeretei, valamint a kutatónak korábbi vagy későbbi nyelvállapotok fonetikai/fonológiai rendszeréről való ismeretei.

További nehézségek származnak abból, hogy egyes történeti dokumentumok a központozást nem használják. Így például a hangsúlyt nem jelölő középkori katalán kéziratokban az első konjugációhoz tartozó igék egyes szám harmadik személyű alakjainak jelen ideje és befejezett múlt ideje formailag egybeesik. Például a *parla* 'beszél' és a véghangsúlyos *parlà* 'beszélt' formák egyaránt *parla* alakban fordulhatnak elő az ilyen szövegekben (l. Bruguera 1981: 31). Így nehéz megállapítani, hogy a kérdéses alakokat múlt idejű vagy pedig jelen idejű igealakként értelmezzük-e. A kontextus bizonyos esetekben sokat segít: egy jelen idejű leírás esetén, ahol a többi, formailag egyértelmű igealak is jelenben szerepel, fel sem merül az, hogy a *parla* alakot befejezett múltként interpretáljuk. Narratív szövegek esetén azonban már nehezebb a dolgunk, hiszen a történeti jelen és a befejezett múlt egyaránt elbeszélő igeidők. Így annak eldöntéséhez, hogy a *parla* alakot egy ilyen szövegben történeti jelenként vagy pedig befejezett múltként értelmezzük-e, az is befolyásolni fogja, hogy milyen ismereteink vannak a történeti jelen használatáról, és hogy a kérdéses alak a diskurzusban hol helyezkedik el: azaz várjuk-e vagy elképzelhetőnek tartjuk-e az adott helyen egy történeti jelen időben álló forma megjelenését. A hasonló esetekben tehát a történeti adat megkonstruálásakor figyelembe kell vennünk számos korábbi történeti kutatás eredményét, a környező igealakokat, a tágabb kontextust, ismereteinket az adott kor írásbeliségéről, elméleti megfontolásokat, valamint a diskurzus szervezésére vonatkozó háttérismereteket is.

4.2 A adatok elméletfüggősége

A Kertész és Rákosi (2012: 174) által felvetett következő probléma, hogy el kell tudnunk dönteni, mely információk relevánsak az adott kutatási kérdés szempontjából, tehát szelektálni kell a potenciális adatok között. A kutatás hátterében mindig valamilyen elmélet áll, így a korpusz mondatainak valamely elmélet segítségével történő elemzése az adatok elméletfüggőségéhez vezet (l. még Lehmann 2004: 183 és Fischer 2007: 17). Amint Kertész és Rákosi (2012: 171) hangsúlyozzák, már egy egyszerű grammatikalitási ítélet mögött is áll valamiféle elmélet, amely vonatkoztatási pontként szolgál a döntés meghozatalakor. Hasonlóképpen, előzetes grammatikai elemzést feltételez annak eldöntése is, hogy egy, a korpuszban talált forma vajon annak a szerkezetnek a példája-e, amellyel a kutató foglalkozni kíván. Ennek megállapítása a történeti kutatásban azért problematikusabb, mert a vizsgált nyelvállapotra nézve a kutató nem

rendelkezik anyanyelvi kompetenciával. Ennek hiányában a kutató gyakran nagyobb mértékben támaszkodik az elmélet sugallta ismeretekre, ami azzal a veszéllyel jár, hogy az elméleti keret sugallta kategóriákat esetleg „belelátja" a történeti előfordulásokba (Fischer 2007: 21). Az elméletvezérelt kutatás példájaként Fischer (2007: 21) egy történeti előfordulás értelmezését mutatja be. A (3)-ban látható szövegrész az angol *habban*/*have(n)* + *to-infinitive* konstrukció egy korai előfordulását tartalmazza.

(3) *For love and joy I had to se her* (Malory, *Wks* Vinaver 1967[8]: 421,13)
(3a) [For love and joy I had] [to se her]
'a szeretet és öröm miatt, amit éreztem, hogy láthatom őt'
(3b) [For love and joy] [I had to se her]
'a szeretet és öröm miatt látnom kellett őt'

A *habban*/*have(n)* + *to-infinitive* szerkezet grammatikalizációs folyamaton ment keresztül, amelynek során modális jelentései is kialakultak. A (3) példa kétféleképpen értelmezhető, ahogyan azt (3a) és (3b) alatt látható. Az elméletvezérelt megközelítésben a korai történeti előfordulásba beleláthatjuk a (3b) alatt megadott értelmezést, mintha abban a ragozott ige már modális jelentéssel bírna. Fischer szerint azonban, ha jobban megnézzük a kontextust, kiderül, hogy a kifejezés még teljes lexikai jelentésében áll a kérdéses szöveghelyen, tehát a (3a) alatt megadott értelemben. A példa mutatja, hogy előzetes ismereteink arról, hogy hasonló jelentésű szerkezetek a grammatikalizációs jelentésváltozás mely állomásain szoktak átmenni a világ nyelveiben ahhoz vezethet, hogy a később kialakuló jelentést már korábbi előfordulásokba is belevetítjük.

4.3 A nyelvi intuíció szerepe és a közvetettség

A korpusz elemzésekor a kutatónak szükségszerűen használnia kell nyelvi intuícióját (Kertész & Rákosi 2012: 173–174). A kutató nyelvi intuíciójának szerepét a korpusz feldolgozásakor jól mutatja Bruguera eljárása (1981: 38), aki a katalán „*anar* + FI" szerkezet I. Jakab krónikájából nyert előfordulásainak jelentését igyekszik megállapítani. Ez a katalán szerkezet befejezett múlt idővé alakult. A folyamat feltárásában különösen fontos annak megállapítása, hogy a korai előfordulásokban az *anar* ige szószerinti 'megy' jelentésében áll-e, vagy pedig már a szerkezet grammatikalizálódott előfordulásával van dolgunk. Bruguera (1981: 38) a szerkezet előfordulásainak jelentését vizsgálva a következő gondolatkísérletet végzi el. A jelentés megállapítására lényegében minimálpárt hoz létre, azaz a vizsgált perifrázis előfordulásait kétféleképpen fordítja le modern katalánra:

(4) va dir
megy.E3.Jel[9] mond.FI 'megy mondani'
(5) va dir
Aux.E3.Jel mond.FI 'mondott'

[8] Vinaver, Eugène (szerk.) 1967: *The Works of Sir Thomas Malory*. Oxford: Clarendon.
[9] A glosszákban használt rövidítések a következők: E3 – egyes szám harmadik személy, Jel – jelen idő, FI – főnévi igenév, Aux – segédige.

Először az *anar* igét főigeként, azaz mozgás jelentéssel fordítja, azaz a (4)-ben látható szerkezetet rendeli hozzá. Másodszor a múlt idő példáinak tekinti az előfordulásokat, azaz a (5)-ben látható szerkezetet rendeli hozzájuk és a modern katalán összetett múlttal fordítja őket. Bruguera szerint a vizsgált szövegrészek a szereplők térbeli elhelyezkedését tekintve olyanok, hogy a szerkezetben megjelenő főnévi igenév által leírt cselekvés térbeli mozgást, helyszínváltást feltételez. Ezt figyelembe véve a szerző megállapítja, hogy a második esetben a fordítás nem vág egybe olyan mértékben az eredetivel, mint az első esetben. Az így keletkezett szövegből ugyanis hiányzik a helyszínek közötti váltás leírása. Ebből azt a következtetést vonja le, hogy a helyes interpretáció a kérdéses esetekben a szószerinti fordítás, azaz a szöveg által képviselt nyelvállapotban a szerkezet grammatikalizációja még nem kezdődött meg.

A kutató nyelvi intuíciójának használata a korpusz felhasználásakor a közvetettség kérdését is felveti. A beszélő nyelvi kompetenciája már a szinkrón kutatásokban sem férhető hozzá közvetlenül, csupán a nyelvhasználó nyelvi megnyilvánulásainak elemzésén keresztül írhatjuk le, a közvetettség kérdése a történeti kutatásban azonban még hangsúlyozottabban jelentkezik. Fischer (2007: 43) megállapítja, hogy a korábbi nyelvállapotokra vonatkozó evidencia minden esetben közvetett (vö. Jacobs & Jucker 1995: 7 és Dömötör 2012: 44). Amennyiben a történeti szövegeket a szöveg szerzője nyelvi kompetenciájának lenyomataként tekintjük, láthatjuk, hogy a hozzáférésben többszörös közvetettséggel van dolgunk, hiszen már maguk a történeti szövegek is a közvetettség számos fokozatát mutathatják. A kutató általában szerkesztett, kiadott verziókkal dolgozik, amelyekben már a kiadó következtetései is tükröződnek. Ilyen értelemben a szöveg kiadójának, átírójának stb. kompetenciája a kutató nyelvész kompetenciájának felhasználása előtt már nyomokat hagy a szövegen. Az így módosult történeti szöveg kerül azután a kutató kezébe, aki a korpuszban található megnyilatkozások értelmezésekor szükségszerűen a saját nyelvi intuíciójára is támaszkodik, ami a közvetettség további fokozatát jelenti. Amikor tehát a történeti szöveg alapján adatokat konstruálunk, minimum három közreműködő nyelvi intuíciójának lenyomatát tartalmazza az adat: a szöveg eredeti szerzőjén túl a másolatot készítő scriptor (ez a kettő természetesen egybe is eshet), a fennmaradt kézirat alapján a megszerkesztett/kiadott verziót készítő személy és a vizsgálatot végző nyelvész nyelvi intuíciója egymásra rétegződik.

A kutató nyelvi intuíciójának használata eredményezi, hogy az így konstruált adat szubjektív lesz, tehát a korpusz semmiképpen sem tekinthető ismeretek objektív forrásának. Consten és Loll (2012) mutatja meg, hogy ez a szubjektivitás hogyan változtatható legalábbis interszubjektivitássá, ami növelheti az adat megbízhatóságát. A szerzőpáros az annotációból származó problémákkal foglalkozik a korpuszadatok kapcsán. Az annotáció a korpusz felhasználásának, az adatok kinyerésének egyik lépése lehet, amelynek során az annotátor maga is elkerülhetetlenül az adat konstruálójává válik, hiszen az elméleti megfontolások az annotációban is visszaköszönnek. Az annotációs kategóriák kialakítása erősen elmélet-vezérelt. Consten és Loll szerint (2012: 712–713) az ebből eredő bizonytalanságokat, problémákat kiszűrni, vagy legalábbis minimalizálni az annotáció átlátszóságának biztosításával lehet. Ez azt jelenti, hogy lehetővé kell tenni, hogy az annotált adatból rekonstruálni lehessen az eredetit. Hogy ez mennyire lehetséges, az az elméleti megközelítéstől is függ: a nyelvi funkcióra vonatkozó hipotézisek esetén a szerzők szerint lehetetlen független, az annotációt vezérlő kritériumot találni, az egyetlen kiindulópont maga az annotálandó anyag és az annotátor konceptuális ismeretei, ami körbenforgáshoz vezethet, és a rekonstruálhatóság kritériumainak nehéz eleget tenni (Consten & Loll 2012: 705, 710). Az említett problémák minimalizálását a szerzők szigorú annotációs irányelvek kidolgozásában látják. Ez a módszer interszubjektívvé teszi az elem-

zést, ugyanis csökken annak valószínűsége, hogy az egyes annotátorok különbözőképpen annotálják a szöveget. Az annotátorok különböző, előre meghatározott heurisztikák alapján dolgozhatnak, így nem nekik kell döntést hozni a versengő lehetőségek között. Ilyen heurisztika lehet például a következő: „ha X a lehetőségek között van, válaszd azt!" (Consten & Loll 2012: 713).

Az eljárás hasonlatos a grammatikalizációs jelentésváltozás vizsgálatában Traugott és Dasher által (2004: 44) javasolt elvhez, hogy amikor csak lehetséges, (azaz amennyiben a kontextusban ennek semmi nem mond ellent), az eredeti lexikai jelentést feltételezzük, vagy legalábbis azt tekintjük kódolt jelentésnek. Érdemes azonban szem előtt tartani, hogy ennek a módszertani elvnek a követése nem vezet el bennünket szükségképpen a szöveg szerzője által szándékolt jelentéshez. Mégis, kétségtelen előnye, hogy interszubjektívvé teszi a leírást.

4.4 A pótkompetencia fogalma

A történeti nyelvészeti kutatás során a nyelvésznek egy korábbi nyelvállapot jelenségeiről kell nyelvi ítéleteket hoznia, amelyre nézve nem rendelkezik megfelelő kompetenciával (vö. Dér 2004: 191–192). Az ún. pótkompetencia kifejlesztésének lehetőségét és korlátait tárgyalja Forgács (1993–1994). A nyelvtörténeti adatról írva hasonló fogalomról tesz említést Dömötör (2012: 45), ő az „elsajátított intuíció" (*acquired intuition*) és „történeti nyelvérzék" kifejezéseket említi. Az ilyenfajta nyelvérzék használatakor a kutató nyelvi intuíciójával kapcsolatos, fentebb tárgyalt problémák még hangsúlyozottabban jelentkeznek.

Forgács (1993–1994: 18) a történeti korpuszról Greule (1982: 72)-re hivatkozva mint ún. zárt korpuszról ír. A zárt korpusz jellemzője (szemben a nyitott korpusszal), hogy az abból kinyerhető információkkal kapcsolatosan esetlegesen felmerülő kérdések tisztázására nem állnak rendelkezésre a korpusz nyelvét anyanyelvi szinten beszélő informátorok, illetve a kutatást végző kutató sem rendelkezik a korpusz nyelvi állapotának megfelelő kompetenciával. A zárt korpuszon alapuló nyelvi leírás megbízhatóságához a kutatónak ún. pótkompetenciát kell kifejlesztenie, hogy maga kísérelhesse meg pótolni az említett hiányt. Belátható, hogy ez sosem érhet teljesen a korabeli nyelvhasználók kompetenciájának nyomába.

A pótkompetencia fogalmához köthető a Fischer (2007: 15) által kívánatosnak tartott rendszer-érzékenység (*sense of system, sense of the grammatical system*) képessége is. A szerző arról írva, hogy nemcsak a vizsgálat szűk értelemben vett tárgyát, hanem egyéb, hasonló nyelvi elemeket is be kell vonnunk a vizsgálatba, hangsúlyozza, hogy fontos, hogy rendelkezzünk egyfajta rendszer-érzékenységgel. A vizsgált korábbi nyelvállapot teljes nyelvtani rendszerére való érzékenység megfeleltethető a pótkompetencia fogalmának.

A pótkompetencia használatának jelensége a fentebb vizsgált közvetettség témaköréhez is kapcsolható. Az, hogy a kutató nyelvész, de a későbbi korban született szövegmásoló sem rendelkeznek már anyanyelvi kompetenciával a szöveg tükrözte nyelvállapotot illetően, a vizsgálni kívánt nyelvállapothoz való hozzáférés nagyobb fokú közvetettségét vonja maga után. Egyrészt kérdéses, hogy vajon az értelmezőnek az értelmezni kívánt szöveg tükrözte nyelvállapotra vonatkozó pótkompetenciája mennyire tekinthető megbízhatónak, másrészt a saját anyanyelvi kompetenciája is visszaköszönhet a szöveg értelmezésében. Vajon mennyire küszöbölhető ki a pótkompetencia és az anyanyelvi kompetencia közötti interferencia? Lényegében erről az interferenciáról beszél Fischer (2007: 18), amikor felhívja a figyelmet arra, hogy történeti szövegek értelmezésekor fennáll annak a veszélye, hogy a mai nyelvállapot szempontjából vesszük szemügyre a régi formát (vö. (3) példa). Ez különösen akkor jellemző, ha a két

alak formailag egybeesik vagy nem sokat változott. A kompetenciák keveredésének veszélye a történeti nyelvérzék fejlesztésével, a vizsgált nyelvállapot teljes rendszerére vonatkozó ismereteink minél magasabb szintre emelésével csökkenthető.

A különböző nyelvállapotokra vonatkozó kompetenciák interferenciája mellett felmerül az is, hogy az írott, illetve a beszélt nyelvre vonatkozó kompetenciánk hogyan befolyásolja egymást. Fischer (2007) megállapítja, hogy ha régebbi korok írott szövegeit későbbi korokéval vetjük össze, az olyan, mintha az almát a körtével hasonlítanánk össze (Fischer 2007: 41). Ugyanis a régebbi korok írott nyelvezete a beszélt nyelvhez közelebb áll. Később, amikor az írásos sztenderd kialakul, az írott nyelv sajátos, azt a beszélt nyelvvel szembeállító jellegzetességeket vesz fel. A grammatikalitási ítéletek azonban gyakran az írott nyelvet veszik alapul, és agrammatikusnak, nem elfogadhatónak ítélünk olyan szerkezeteket, jelenségeket, amelyek a beszélt nyelvben előfordulnak. Így Fischer felhívja a figyelmet arra (Biber & Finegan 1994-re hivatkozva), hogy ha szövegek egy homogén csoportját akarjuk egymással összevetni, nem a műfajt kell figyelembe venni, hanem érdemes inkább nyelvi mutatókra támaszkodni (mint pl. passzív szerkezetek, alárendelések vagy első/második személyű névmások jelenlétének mértéke a szövegben) (Fischer 2007: 13).

4.5 A pozitív és negatív evidencia kérdése

Végül, a korpuszhasználattal kapcsolatban felmerülő utolsó probléma a pozitív és negatív evidencia kérdése. Nagyon leegyszerűsítve, evidencián Kertész és Rákosi (2012: 178) olyan adatot ért, „amelynek funkciója az, hogy *hozzájáruljon a rivális hipotézisek plauzibilitásának megítéléséhez és egymással való összevetéséhez*" (kiemelés az eredetiben).[10]

4.5.1 Pozitív evidencia

Pozitív evidencia esetén egy, a kutató által vizsgált jelenséget képviselő szerkezet előfordulása megtalálható a korpuszban. Első megközelítésben ez plauzibilissé teszi azt a kijelentést, hogy a kérdéses nyelvi konstrukció grammatikus a vizsgált nyelvben (történeti kutatásról írva: nyelvállapotban). Kertész és Rákosi (2012: 19–20, 69 és 173–175) felhívják a figyelmet arra, hogy ez nem feltétlenül van így, hiszen a korpuszban előfordulhatnak elírások, hibás alakok (vö. a 4.1. pontban tárgyalt problémával). Hozzá kell tenni, hogy az sem bizonyos, hogy a kérdéses forma tényleg a vizsgált jelenség példája-e, ennek eldöntése a kutató szubjektív ítéletén múlik (vö. a 4.2. pontban tárgyalt problémával).

Mivel, ahogy fentebb említettem, egyazon történeti szöveg különböző részei akár különböző nyelvállapotokat is tükrözhetnek, a pozitív evidenciával kapcsolatos problémák a történeti nyelvészetben még hangsúlyosabban jelentkeznek, hiszen ahhoz, hogy plauzibilitási értéket rendelhessünk ahhoz a kijelentéshez, hogy a szövegben szereplő nyelvi konstrukció grammatikus a vizsgált nyelvállapotban, tudnunk kell, hogy a kérdéses szövegrész mely nyelvállapotot tükrözi. Lényeges tehát, hogy a kutató el tudja különíteni a szöveg ténylegesen a vizsgált nyelvállapotot tükröző részeitől az esetleges elírásokat, későbbi betoldásokat, korábbi vagy későbbi nyelvállapotot tükröző szövegrészeket, hiszen az ezekben talált előforduláson alapuló adat nem tekinthető pozitív evidenciának a kérdéses nyelvállapot vizsgált jelenségére nézve.

[10] „(evidence is a datum) whose function is to *contribute to the judgement and comparison of the plausibility of rival hypotheses*" Az evidencia fogalmához l. még Kertész & Rákosi (2008d).

Nagy C. Katalin:
Adatforrások integrációja a történeti nyelvészetben: a nyelvtörténeti korpuszadat fogalmáról
Argumentum 9 (2013), 56-78
Debreceni Egyetemi Kiadó

4.5.2 Negatív evidencia

Még nehezebb a negatív evidencia értelmezése. Negatív evidencia esetén a korpusz nem tartalmazza a vizsgált nyelvi konstrukció egyetlen előfordulását sem, ami első megközelítésben implauzibilissé teszi azt a kijelentést, hogy a nyelvi konstrukció grammatikus az adott nyelvben (nyelvállapotban) (l. Kertész & Rákosi 2012: 19–20 és 69). Egy adott szerkezet hiánya a korpuszban azonban nem értelmezhető automatikusan negatív evidenciaként. Előfordulhat, sőt bizonyos esetekben megalapozottnak tűnik feltételezni, hogy a kérdéses forma bővebb korpuszban megtalálható lenne, és az is lehet, hogy elő is fordul, csupán elkerülte a figyelmünket.

Vizsgáljuk meg ezt a problémát Bybee (2005 [2003]: 605–614) egy kutatásán, amelyben az angol *can* ige segédigévé válását vizsgálja óangol és középangol korpusz alapján. A típusgyakoriság vizsgálatakor intellektuális állapot vagy tevékenység igék, kommunikációs igék és képesség igék között tesz különbséget, hiszen ezek azok a főnévi igenév típusok, amelyek kezdetben tipikusan megjelennek a *can* igével. 300 előfordulás alapján a középangolra nézve a következő számadatokat kapja a vizsgált szemantikai csoportokat illetően:

1. Intellektuális állapot vagy tevékenység igék: 52 példány, 18 típus
   - magas példánygyakoriság: *see* 12, *deem* 6, *understand* 6, *espy* 5
   - típusgyakoriság: 18 különböző ige
2. Kommunikációs igék: 102 példány, 31 típus
   - magas példánygyakoriság: *tell* 30, *say* 29, *devyce* 8
   - típusgyakoriság: 31 különböző ige
3. Képesség igék („tudni, hogyan”): 26 példány, 18 típus

Bybee a fenti, középangolra vonatkozó arányokat az óangolra kapott adatokkal veti össze és megállapítja, hogy bár ezek a szemantikai csoportok az óangolban is felfedezhetőek, a középangolban minden csoportban több főnévi igenév szerepel, és egyes kombinációk példánygyakorisága igen megnőtt az első két csoportban. Továbbá újabb főnévi igenevek is megjelennek a *can* segédigével. Bybee szerint mindkét fajta gyakoriságnövekedés hozzájárul a vizsgált nyelvi elem jelentésének kifakulásához. A leírásból jól látszik, hogy Bybee jelentőséget tulajdonít annak, hogy pl. a kommunikációs igék csoportjában 31 különböző igére talált előfordulást a középangol korpuszban, míg az óangol korpuszban jóval kevesebbre. A középangol korpuszban szereplő, de az óangol korpuszból hiányzó igék tehát negatív evidenciaként funkcionálnak az óangol nyelvállapot leírásakor, hiszen hiányuk mutatja, hogy a vizsgált szerkezet még nem volt általánosan használatos az óangol nyelvállapotban. Ha azonban a középangol korpuszban nem találna előfordulást egy az általa felsorolt igékkel jelentésében rokon igére, annak feltehetően nem tulajdonítana ilyen jelentőséget. Az a megfigyelés ugyanis, hogy kommunikációs igék típus- és példánygyakorisága a középangol korpuszban már igen magas, valószínűsíti, hogy egy további kommunikációs ige hiánya pusztán a véletlennek köszönhető, és a korpusz bővítése esetén arra is találnánk előfordulást. Hasonlóképpen, ha egy modern angol szövegeket tartalmazó korpuszban nem találnánk például a – már a középangol korpuszban is többször megjelenő – *can say* kombináció egyetlen előfordulását sem, azt nem értelmeznénk negatív evidenciaként a kifejezés agrammatikussága mellett. Csupán a véletlennek tulajdonítanánk, és a korpusz bővítésével a kérdéses szerkezet előfordulását várnánk.

Tekintsünk egy másik példát! A katalán „*anar* + FI” szerkezet grammatikalizációs jelentésváltozás eredményeképpen befejezett múlttá vált. A modern katalánban ebben a szerkezetben az *anar* ige már nem őrzi eredeti, mozgásra utaló jelentését (’megy’), hanem pusztán segédigeként funkcionál. A szerkezet jelentésváltozásának korai szakaszában megjelent a szerkezet prepozíciós változata („*anar a* + FI”) is azoknak az eseteknek az elkülönítésére, amikor a nyelvhasználó az *anar* igét szószerinti ’megy’ jelentésben akarta használni. A prepozíciós szerkezet jelenléte ilyen értelemben tükrözi a prepozíció nélküli szerkezet grammatikalizációjának az előrehaladását. Az erre vonatkozó kutatásban tehát jelentőséget tulajdonítunk annak, ha az *a* prepozíciós előfordulások egy adott nyelvállapotot reprezentáló korpuszban még nem fordulnak elő: hiányukat úgy értelmezzük, hogy a prepozíció nélküli szerkezet grammatikalizációja még nem kezdődött meg. Elképzelhető azonban, hogy a korpusz bővítésével olyan szöveget is lelhetünk, amelyben a kérdéses alak előfordul, ekkor módosítanunk kell a szerkezet grammatikalizációjának datálását. Az *a* prepozíciós szerkezet példáinak hiányát bizonyos esetekben tehát negatív evidenciaként kezelhetjük. Ezzel szemben nem tekintjük negatív evidenciának, ha egy későbbi nyelvállapotot reprezentáló korpuszban, amelyben már a „’megy’ + FI” szerkezet *a* prepozíciós előfordulásai is megtalálhatóak, egy bizonyos főnévi igenévvel való kombinációnak nem leljük *a* prepozíciós előfordulását. Mivel a más főnévi igenevekkel való *a* prepozíciós előfordulások arra utalnak, hogy a szerkezet mint típus már megjelent, azt feltételezzük, hogy bővebb korpuszban a hiányzó főnévi igenévvel való példája is előfordulhatna.

Egy vizsgált szerkezet vagy jelenség hiánya egy szövegben egyes esetekben műfaji sajátosságoknak is betudható. Colon (1978a [1959]) mintegy száz szöveget vizsgál meg, amelyeknek csupán a felében találja meg a katalán „*anar* + FI” szerkezet előfordulásait. Colon (1978a [1959]: 120) ezt a mennyiségű szöveget már elégnek tartja ahhoz, hogy bizonyos összefüggéseket állapítson meg arra nézve, hogy az adott korban a vizsgált szerkezetet mely regiszterekben használták és melyekben nem. Az, hogy egyes szövegtípusokból hiányzik a szerkezet, negatív evidenciaként funkcionál annak elemzésekor, hogy a szerkezet mely regiszterekben volt elfogadható a vizsgált nyelvállapotban. Természetesen ez a megállapítás csak akkor érvényes, ha minden szövegtípusból elegendő mennyiségű[11] szöveg áll rendelkezésünkre. Egy vizsgált szerkezet hiánya egyes esetekben a szöveg rövidségének vagy stílusának is betudható. Természetesen minél nagyobb a korpuszunk, annál nagyobb súllyal esik a latba a negatív evidencia. A korpusz bővítésének azonban, különösen a történeti nyelvészetben, de nemcsak ott, megvannak a korlátai, így negatív evidencia alapján soha nem tekinthetünk egy hipotézist teljesen bizonyosnak.

Lehetetlen tehát teljes bizonyossággal eldönteni, hogy egy forma hiánya az adott nyelvállapot grammatikai jellemzőinek, a szöveg műfajának, a puszta véletlennek, vagy pedig a szöveg írója egyedi stílusának, ízlésének tudható be. Ez utóbbinak érdekes példáját írja le Soldevila (1963–1968) a katalán „*anar* + FI” szerkezet írásbeli használatával kapcsolatban. A kérdéses igeidő, bár mára a beszélt nyelvben a katalán szinte minden változatában elterjedt, irodalmi vagy tudományos írásokban még nem szorította ki teljesen az egyszerű befejezett múltat. Egyik vagy másik befejezett múlt használata sok esetben a szöveg szerzőjének egyedi stílusán is múlik. Soldevila leírja, hogy ő maga, szemben más szerzőkkel, stilisztikai célzattal alkal-

[11] Újabb problémát jelent annak megállapítása, hogy mennyi az elegendő. Azt mindenesetre elfogadhatjuk, hogy a korpusz méretének növelésével a különböző szövegek összehasonlítása alapján levont következtetések plauzibilitása is növekszik.

mazza tudományos írásaiban is a szerkezetet, hogy ne legyen unalmas olvasmány. Soldevila megfigyeli, hogy a középkori, Muntaner által írt krónikában az írás vége felé az „*anar* + FI" előfordulásai elszaporodnak, amit ő – saját maga nyelvhasználatát alapul véve – annak tud be, hogy a szerző „fellelkesült" a szerkezet használatán, azaz a szerző egyéni stílusának tulajdonítja a szerkezet egyazon kontextusban való halmozott használatát.

Összefoglalva, sem egy adott forma hiánya, sem a korpuszban való megléte nem segít annak teljes bizonyossággal való eldöntésében, hogy a kérdéses konstrukció vajon megtalálható-e a korpusz által reprezentált nyelvállapotban vagy nem. Mindezen megfontolások alapján jó okunk van rá, hogy a korpuszt, és így természetesen a történeti korpuszt se tekintsük teljes mértékben megbízható forrásnak. Mivel a korpuszok csupán többé-kevésbé megbízható forrásnak tekinthetők, és egy-egy nyelvi jelenség magyarázatakor más adatforrásokkal kell őket kiegészítenünk (l. Kertész & Rákosi 2012: 173–175, valamint Kertész & Rákosi 2008d), a korpuszadatok soha nem tekinthetőek teljesen bizonyos, csupán plauzibilis kijelentéseknek. Fischer (2007: 14) is hasonló konklúzióra jut, amikor kijelenti, hogy mást nem tehetünk, mint hogy a kutatásban világossá tesszük, hogy mely szövegeket választottuk és miért, és jelezzük a történeti szövegek jellegének betudható bizonytalanságokat.

5 A történeti adat forrásának integrált természetéről

Mindeddig úgy beszéltünk a korpuszadatokról mintha azok pusztán a korpuszon mint adatforráson alapuló adatok lennének. A korpuszadatokkal kapcsolatos 3. problémából kiindulva (l. 4.3. pont) azonban további megfigyeléseket tehetünk. A korpuszon alapuló adatok egyidejűleg a kutató nyelvész intuícióján alapuló adatok is, hiszen a korpusz feldolgozása során elengedhetetlenül szükség van a befogadó nyelvi intuíciójára. Gondolhatunk itt a grammatikalitási ítéletekre is, hiszen a történeti szövegek feldolgozása során a kutatónak először is el kell döntenie, hogy a korpuszban fellelhető nyelvi forma grammatikus alak-e, vagy esetleg elírásról van-e szó. Ez még a nyelvi intuíció használatának viszonylag problémamentesebb formája, bár már itt is lehet különbség egyes kutatók nyelvi ítéletei között. Ha azonban a jelentés szintjére, vagy annak is a pragmatika által vizsgált nehezebben megközelíthető szintjeire lépünk, a véleménykülönbségek feltehetően még kiélezettebbek lesznek.

Ha a történeti adatot a fentebb bemutatott értelemben, a p-modell adatfogalmát alkalmazva fogjuk fel, akkor azt a történeti szövegben talált nyelvi elem jelenlétére vagy annak valamely tulajdonságára vonatkozó kijelentésnek kell tekintenünk, amely plauzibilitási értékkel is rendelkezik. A nyelvész intuícióját mint forrást ennek megfelelően a történeti adat információtartalmának és plauzibilitási értékének vonatkozásában is megvizsgálhatjuk. Láthattuk, hogy a nyelvi intuíció már a kijelentés megfogalmazásában is szerepet kap. Ugyanakkor az adat plauzibilitását is befolyásolja, hogy a kutató intuícióját milyen mértékben tekintjük megbízható forrásnak. A történeti korpuszadat esetében a pótkompetencia kérdése is felmerült. Megbízhatóbbnak kell-e tekintenünk az anyanyelvi kompetenciát mint forrást a pótkompetenciánál? Erre a kérdésre azért is nehéz válaszolni, mert a kettő működését a történeti szövegek elemzésében esetenként szinte lehetetlen elkülöníteni.

Bruguera (1981) fentebb (a 4.3. pontban) bemutatott eljárása jó példa arra, hogyan játszik szerepet a nyelvész intuíciója a korpuszban lelt előfordulások értelmezésekor. Ennek az eljárásnak az alkalmazásával Bruguera a következő adathoz jut: „Plauzibilis, hogy a középkori katalán „*anar* + FI" szerkezet I. Jakab krónikájában található előfordulásaiban az *anar* ige

lexikális, 'megy' jelentésében áll." Ennek a történeti adatnak a forrása tehát nemcsak maga a történeti szöveg lesz, hanem a kutató nyelvi intuíciója is.

A kutató nyelvi intuícióján kívül egyéb források is szóba jöhetnek a történeti korpuszadat megkonstruálásakor. A történeti adat tehát integrált forrásból származó adat. Tekintsünk egy-két esetet arra vonatkozóan, hogy hogyan épül fel ez az integrált forrás, azaz milyen információk egyesülnek azokban a plauzibilis kijelentésekben, amelyeket hagyományosan korpuszadatnak, esetünkben történeti korpuszadatnak hívunk.

a. példa

A történeti szövegek interpretációjában nehézség származhat már maguknak a szövegeknek a jellegéből is: az írásbeliség kialakulásának kezdetén ugyanis még nem volt egységes írott norma, a központozás hiányozhat, a nyelvi elemek számos írásmóddal jelenhetnek meg még egyazon szövegen belül is. Így például egy középkori spanyol nyelvű korpusz szövegeiben a 'fogad' jelentésű ige többféle írásmóddal fordulhat elő, úgy mint pl. *recebir, reçebir* vagy *rresçebir***.** Ahhoz, hogy a kutató megállapítsa, hogy mindezek egyazon főnévi igenév előfordulásainak tekintendőek, a nyelvi intuícióját is használja. Ráadásul, míg a különböző írásmódok egyes esetekben pusztán az írásos norma hiányának tudhatóak be, máskor lehetnek szimpla elírások, megint máskor esetleg valamely hangváltozásról tanúskodnak. Hogy el tudjuk dönteni, melyikről van nagyobb valószínűséggel szó, – már ebben a viszonylag egyszerű esetben is –, a történeti korpuszadat megkonstruálásakor a következőket is figyelembe kell vennünk: a kutató nyelvi intuíciója, további előfordulásokra vonatkozó információk, a kérdéses hangok írásos reprezentációi más szavakban, a fonetikai/fonológiai változás korábban feltárt törvényszerűségei, a „hiba" vagy írásmódbeli különbség helye, a fonetikai/fonológiai rendszer az adott nyelvállapotban, valamint a fonetikai/fonológiai rendszer a korábbi nyelvállapot(ok)-ban. Egy történeti fonológus számára a fenti előfordulásokból csak akkor lesz a kutatásában felhasználható adat, ha mindezekre választ ad.

b. példa

Tekintsünk egy másik példát. A középkori katalánban formai egyezés volt mindhárom konjugációhoz tartozó igék esetében a többes szám első személyű alakok jelen és befejezett múlt ideje között. Így nehéz megállapítani, hogy a kérdéses alakokat történeti jelenként vagy pedig befejezett múltként értelmezzük-e (Bruguera 1981: 31). Így például a katalán *anar* 'megy' ige többes szám első személyű alakja (*anam*) mind jelen, mind befejezett múlt időként interpretálható ('megyünk/mentünk') a középkori nyelvállapotban (l. Juge 2006: 320). Hogy hogyan értelmezzük az egyes szöveghelyeken, az alapjaiban befolyásolja a katalán „*anar* + FI" szerkezet történetéről alkotott hipotéziseket. A kérdéses formákat Coromines (1980–1991) történeti jelenként, Bruguera (1981) pedig befejezett múltként interpretálja I. Jakab krónikájában. Bruguera úgy érvel, hogy ezeken a kétértelmű formákon kívül nincs más jelen idejű alak a szövegben.

Ebben az esetben tehát a kérdéses formák értelmezéséhez, és így a történeti adat megkonstruálásához figyelembe kell vennünk a következőket is: számos korábbi történeti kutatás eredménye, több igealak (többi konjugáció igéi), a tágabb kontextus, környező igealakok morfológiájának figyelembevétele, ismereteink az adott kor írásbeliségéről.

c. példa
A katalán befejezett múlt („*anar* + FI") kialakulásának vizsgálatakor az *anar* 'megy' igével alkotott szerkezet prepozíció nélküli és prepozíciós verziójának megoszlását is vizsgálnunk kell. Az egyik középkori katalán krónika, Desclot krónikája az „*anar* + FI" szerkezet két *a* prepozíciós előfordulását is tartalmazza. Mindkettő címben szerepel azonban, és mivel tudjuk, hogy a címeket gyakran később illesztették a szöveghez, ezeket az előfordulásokat egy későbbi nyelvállapothoz tartozónak kell tekintenünk (l. Pérez Saldanya 1996: 74). Az „*anar* 'megy' + FI" és az „*anar* 'megy' *a* + FI" szerkezetek megoszlásának vizsgálatakor tehát az *a* prepozíciós előfordulások státusának megítéléséhez figyelembe kell vennünk: az előfordulások helyét a szövegben, több kézirat összevetéséből származó információkat, valamint enciklopédikus információkat a történeti szövegről.

Kertész és Rákosi (2012: 169–170) arra is felhívják a figyelmet, hogy az a forrás, amely alapján plauzibilitási értéket rendelünk az adathoz, gyakran összetettebb természetű, mint amit általában az adat forrásának szoktunk tekinteni. Egyrészt az adat több információt tartalmaz, mint ami pusztán magából a korpuszból kinyerhető, hiszen már egyfajta nyelvészeti elemzés eredményét is tartalmazza. Sőt, még számos egyéb információt is, ahogyan a fenti példák mutatják. Másrészt, ha adaton csak annyit értünk, hogy egy adott mondat vagy szósor egy adott nyelvhez tartozik vagy megtalálható egy korpuszban, akkor ahhoz, hogy ezzel a nyelvészeti elméletalkotás folyamatában bármit is kezdeni tudjunk, be kell vonnunk egy másik adatot is, amelyik ennek a mondatnak a szerkezetét ragadja meg (Kertész & Rákosi 2012: 169–170).

Összefoglalva, a korpuszadatok a fentiek alapján integrált adatforrásból származó adatoknak tekinthetők: a korpusz mint adatforrás mellett további adatforrásként szerepel a kutatónak a korabeli és a vizsgált nyelvállapotra vonatkozó nyelvi intuíciója, a korábbi kutatások eredményei, valamint az elmélet. A korpuszadat tehát csak részben korpuszon alapuló adatnak, vagy egy korpuszbeli előfordulásból kiinduló, az alapján konstruált adatnak tekinthető.

A történeti korpuszadat megkonstruálásakor a korpuszbeli előfordulásokon túl tehát további információkat is számításba veszünk, amelyek (többek között) a következőkre vonatkozhatnak: a vizsgált nyelvi forma további előfordulásai, az előfordulás helye, a tágabb kontextus, további hasonló kontextusok, a teljes nyelvállapot, korábbi és későbbi nyelvállapot(ok), a szinkrón nyelvállapot, maga a történeti szöveg és esetleges további kéziratai, fordításai, az adott kor írásbelisége, valamint más nyelvekre vonatkozó, tipológiai információk.

6 A korpuszon alapuló adatok a történeti kutatásban

A korpuszadatok szerkezetének leírásakor Kertész és Rákosi (2012: 173) alapján a korpuszadatoknak olyan szerkezetet tulajdonítottunk, amelyben a plauzibilitási érték a következő kijelentések valamelyikéhez kapcsolódik:
*„**az X nyelvi jelenséget tartalmazó** M megnyilatkozás megtalálható K korpuszban"*
*„az X szerkezet beazonosítható K korpuszban és **rendelkezik Y tulajdonsággal**"*
„a P példában látható mondat K korpusz része" és *„a P példában látható mondat **az SZ szerkezet egy előfordulását tartalmazza**"*.

A fenti kijelentések megfogalmazásában vastagon kiemelt részek már önmagukban mutatják a korpuszon alapuló adatok integrált természetét, hiszen annak megállapítása, hogy a korpuszban beazonosított nyelvi elemek sorozata milyen nyelvi jelenséget képvisel, milyen tulajdonsággal rendelkezik, vagy mely szerkezet előfordulásának tekinthető, már a korpuszon kí-

vül további források bevonását jelenti. A történeti adat tehát direkt források együtteséből származó adat (l. Kertész & Rákosi 2012: 173).

Ha tehát nyelvtörténeti adaton olyan plauzibilis kijelentést értünk, amelynek plauzibilitási értéke direkt forrásból származik, a korpusz mint történeti adatforrás további adatforrásokkal egészül ki, amelyek a következők lehetnek:

- a kutató nyelvi intuíciója
- a kutató történeti nyelvérzéke
- korábbi kutatások eredményei
- elmélet
- következtetések.

Az így megkonstruált történeti korpuszadat plauzibilitási értéke azután a kutatás folyamatába bekerülve módosulhat, azt más forrásokból származó adatok plauzibilitási értéke is befolyásolhatja. A fenti megfontolások alapján összefoglalóan elmondható, hogy a történeti adat csupán részben korpuszon alapuló adat, amely a kutatás és érvelés folyamatában további adatokkal együttesen jelenik meg és csak így funkcionálhat. Ezek a további adatok a következőkre vonatkozhatnak:

- több előfordulás
- az előfordulás helye
- tágabb kontextus
- további releváns kontextusok
- a teljes nyelvállapot
- korábbi és későbbi nyelvállapotok
- maga a történeti szöveg, az adott kor írásbelisége
- több kézirat, fordítások
- más nyelvekre vonatkozó, tipológiai adatok

Összefoglalásképpen elmondhatjuk, hogy a plauzibilis kijelentésekként felfogott korpuszadatok direkt forrása első megközelítésben a korpusz. Mivel azonban a korpuszadatok esetén minden esetben hivatkozunk egyéb direkt forrásokra, és a korpuszadat plauzibilitási értékét a továbbiakban más forrásokból származó adatok plauzibilitási értéke is befolyásolja, a korpuszadat integrált forrásból származó adatnak tekinthető. Következésképpen a korpuszadat terminus helyett találóbb lenne a „részben korpuszon alapuló adat” vagy „korpusz felhasználásával konstruált adat” kifejezést használni.

7 Integráció három szinten

A bevezetésben felvetett integráció fogalma a tanulmányban felvetett problémák kontextusában tehát legalább háromféleképp értelmezhető. Először, a korpuszadat csupán „részben korpuszon alapuló adat”, ilyen értelemben a történeti adat megkonstruálása során az adatforrások integrációjára van szükség. A kutatói nyelvi intuíció állandó jelenléte a kutatás során megkérdőjelezi, hogy léteznek-e egyáltalán „tiszta adatok”, azaz csupán egyetlen forrásból származó adatok. Ilyennek pusztán a nyelvi intuíción alapuló adat tekinthető. Mivel a nyelvi intuíció a nyelvtörténeti kutatásban is szükségképpen jelen van, a nyelvtörténeti korpuszadat mindig integrált forrásból származó adat. Bonyolítja a helyzetet, hogy a történeti kutatásban maga a nyelvi intuíció mint forrás is integrált természetű, hiszen a korábbi nyelvállapotok anyanyelvi

beszélőinek kompetenciája közvetlenül nem hozzáférhető, maga a kutató pedig pusztán pótkompetenciával, történeti nyelvérzékkel rendelkezhet a korábbi nyelvállapotot illetően. A pótkompetencia működését ráadásul befolyásolhatja a kutató szinkrón nyelvállapotra vonatkozó, illetve anyanyelvére vonatkozó nyelvi intuíciója is.

Másodszor, integrációról beszélhetünk az adat „használata", azaz az érvelés folyamán is. Ez a folyamat különböző (integrált) forrásokból származó adatok integrációját foglalja magában, hiszen a történeti adat csak így funkcionálhat.

Harmadszor, az adatforrások integrációja figyelhető meg az adat szerkezetén belül is. A források integrációját a történeti korpuszadatokkal kapcsolatban is két szinten lehet megvizsgálni: a kijelentés információtartalmára és a plauzibilitási értékére vonatkozóan. Az adat szerkezete ugyanis összetett, és azon belül a kijelentés információtartalma és plauzibilitási értéke más-más adatforrásból is származhat. A kijelentés információtartalma származhat például a korpusz, valamely elmélet és a kutató nyelvi intuíciójának mint forrásoknak az integrációjából, míg a plauzibilitási érték a kutató nyelvi intuíciójából.

Végül felmerül a kérdés, hogy akkor vajon hogyan érdemes a különféle adatforrásokat integrálni egymással. Vannak-e szerencsésebb és kevésbé szerencsés integrációk? Vajon mindig előnyösebb-e minél több adatforrás integrálása?

Geluykens és Kraft (2008) felhívják a figyelmet arra, hogy több adatforrás együttes használata nem feltétlenül garantálja az eredmények nagyobb megbízhatóságát. A szerzők a pragmatikában használt egyes adatgyűjtési eljárásokat vizsgálva megállapítják, hogy a kiváltott adatok (*controlled elicitation data*) (diskurzus kiegészítéses tesztek, szerepjátékok) használata nem vezet megbízható eredményhez, mert soha nem lehetünk biztosak abban, hogy a talált összefüggés nem az adatgyűjtés módjának tudható-e be. Véleményük szerint hiába kombinálunk egymással diskurzus kiegészítéses tesztek és szerepjátékok alapján nyert adatokat, az nem vezet az eredmények nagyobb megbízhatóságához, hiszen mindkettő az adatgyűjtés „mesterséges" módja (Geluykens & Kraft 2008: 94). A spontán diskurzusból vett adatok használatát ezért elkerülhetetlenül szükségesnek tartják.

A szerzők sorra veszik azokat a tanulmányokat, amelyek különböző pragmatikai adatgyűjtési módszereket hasonlítanak össze, úgymint diskurzus kiegészítéses teszt, szerepjáték, feleletválasztós teszt. Bár az ezek által produkált adatok hasonlóak lehetnek, két alapvető különbséget figyelhetünk meg. Az adatgyűjtés ilyen módszerei másként hatnak az anyanyelvi és a nem anyanyelvi beszélőkre, így a kapott eredmények összehasonlíthatósága kétséges. Másrészt az így gyűjtött megnyilatkozások eltérnek a természetes nyelvhasználatban megfigyelt viselkedéstől, ami már problematikusabb. A bemutatott tanulmányok felhívják a figyelmet arra, hogy nem elegendő pusztán egy módszer alkalmazása, arra azonban nem adnak választ, hogyan lehet sikeresen integrálni a különböző adattípusokat, és arra sem, hogy az autentikus és a kontrollált adatok kombinálása kívánatos-e. Geluykens és Kraft szerint az mindenesetre egyértelműen megállapítható, hogy ha a valós nyelvhasználatról akarunk bármiféle megállapítást tenni, akkor nélkülözhetetlen az autentikus nyelvhasználat vizsgálata. Az előidézett adatok esetén a nem-autentikus, nem-interaktív jelleg ugyanis befolyásolja a nyelvi megformálást is. Mindezek a megfontolások elvezetnek oda, hogy egyáltalán mely adatforrásokat *érdemes* kombinálni, vajon két adatforrás kombinálása mikor vezet az eredmények nagyobb megbízha-

tóságához. A szerzők azt sugallják, akkor, ha olyan adatforrást vonunk be, amely valamilyen módon „kiküszöböli" a másik adatforrás hibáját[12] (Geluykens & Kraft 2008: 94).

A szerzőpáros által felvetett kérdéseket, valamint azt, hogy maga a kutatási kérdés mennyiben határozza meg, hogy mely adattípusokat érdemes egymással kombinálni, a történeti kutatásra vonatkoztatva is érdemes lenne megvizsgálni, ez azonban már meghaladja jelen tanulmány lehetőségeit.

8 Összefoglalás

A történeti korpuszból nyert adatok felhasználása, megbízhatósága számos kérdést, problémát vet fel, amelyek adódhatnak magának a szövegnek, de az adatkinyerés folyamatának jellegzetességeiből is. Ugyanakkor a történeti kutatás gyakorlata arra utal, hogy a történeti dokumentumokban fellelt előfordulások önmagukban nem funkcionálhatnak adatként a kutatás folyamatában. Ennek ellenére a nyelvtörténeti adat fogalmát a szakirodalomban általánosan egy vizsgált nyelvi elem/jelenség történeti dokumentumokban fellelt előfordulásaira vonatkoztatva használják. Tanulmányomban a nyelvtörténeti adatokról metodológiai szempontból gondolkodva tettem fel a kérdést: hogyan értelmezhető a 'nyelvtörténeti adat' fogalma, mire épülnek a diakrón kutatás hipotézisei, és a történeti dokumentumokban fellelt előfordulások hogyan illeszkednek a kutatás folyamatába.

Írásomban a nyelvészeti elméletalkotás egy modelljére, a p-modellre (Kertész & Rákosi 2012) alapozva, és a nyelvtörténeti kutatás gyakorlatából vett példák segítségével mutattam meg, hogy a nyelvtörténeti korpuszadat integrált adatforrásból származó plauzibilis kijelentésként értelmezhető. A történeti korpuszadat ilyen felfogásával megragadhatóvá válnak a történeti adat felhasználásával kapcsolatos bizonytalanságok, valamint a történeti adat forrásának integrált természete. Hogy a gyakorlati kutatásban az integráció mely formája célravezető, és hogy az integráció milyen módja vezet plauzibilisebb eredményekhez, az további vizsgálat tárgya lehet.

Irodalom

Biber, Douglas & Finegan, Edward (eds.) (1994): *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press.

Bruguera, Jordi (1981): La locució prepositiva „de part", el present històric e el perfet perifràstic en la Crònica de Jaume I. In: *Estudis de llengua i literatura catalanes III, Miscel·lània Pere Bohigas I.* Barcelona: Publicacions de l'Abadia de Montserrat, 27–42.

Bybee, Joan L. (2005[2003]): Mechanisms of Change in Grammaticization: The Role of Frequency. In: Joseph, Brian D. & Janda, Richard D. (eds.): *The Handbook of Historical Linguistics*. Oxford: Blackwell, 602–623.

[12] A nagyobb megbízhatóság Geluykens és Kraft (2008: 94) szerint nem jár automatikusan együtt több adatforrás együttes alkalmazásával, hacsaknem az egyik „autentikusabb" adatokat eredményez („*unless at least one data set consists of more authentic material*"), ami alatt a szerzők a spontán diskurzuson alapuló adatokat értik.

Colon, Germà (1978a[1959]): El perfet perifràstic català "va + infinitiu". In: Colon, Germà (ed.): *La llengua catalana en els seus textos.* Volum 2. Barcelona: Curial, 119–130.

Consten, Manfred & Loll, Annegret (2012): Circularity effects in corpus studies – why annotations sometimes go round in circles. *Language Sciences* 34, 702–714.

Coromines, Joan 1980–1991: *Diccionari etimològic i complementari de la llengua catalana*. Barcelona: Curial Edicions Catalanes.

Dér, Csilla Ilona (2004): Határok nélkül: a grammatikalizáció státusáról. *Nyelvtudományi Közlemények* 101, 182–194.

Dömötör, Adrienne (2012): A nyelvtörténeti adat: elvek, gyakorlat, lehetőségek. *Magyar Nyelv*, 39–51.

Fischer, Olga 2004: What counts as evidence in historical linguistics? *Studies in Language* 28/3, 710–740.

Fischer, Olga 2007: *Morphosyntactic change: Functional and formal perspectives.* Oxford: Oxford University Press.

Fitzmaurice, Susan & Taavitsainen, Irma (2007): Historical pragmatics: What it is and how to do it. In: Fitzmaurice, Susan M. & Taavitsainen, Irma (eds.): *Methods in Historical Pragmatics.* [Topics in English Linguistics 52]. Berlin & New York: Mouton de Gruyter, 11–36.

Forgács, Tamás (1993–1994): Zárt korpuszok és pótkompetencia. *Néprajz és Nyelvtudomány* 35, 17–23.

Francis, W. Nelson (1992): Language corpora B. C. In: Svartvik, Jan (ed.): *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 17–32.

Geluykens, Ronald & Kraft, Bettina (2008): The use(fulness) of corpus research in cross-cultural pragmatics: Complaining in intercultural service encounters. In: Romero-Trillo, Jesús (ed.): *Pragmatics and Corpus Linguistics. A mutualistic entente.* Berlin & New York: Mouton de Gruyter, 93–117.

Greule, Albrecht (1982): *Valenz, Satz und Text. Syntaktische Untersuchungen zum Evangelienbuch Otfrieds von Weißenburg auf der Grundlage des Codex Vindobonensis.* München: Fink.

Heine, Bernd (2005[2003]): Grammaticalization. In: Joseph, Brian D. & Janda, Richard D. (eds.): *The Handbook of Historical Linguistics*. Oxford: Blackwell, 575–601.

Jacobs, Andreas & Andreas H. Jucker (1995): The historical perspective in pragmatics. In: Jucker, Andreas H. (ed.): *Historical pragmatics*. Amsterdam, Philadelphia: Benjamins, 3–27.

Jucker, Andreas H. (2009): Speech act research between armchair, field and laboratory. The case of compliments. *Journal of Pragmatics* 41, 1611–1635.

Juge, Matthew L. (2006): Morphological factors in the grammaticalisation of the Catalan "go" past. *Diachronica* 23/2, 313–339.

Kepser, Stefan & Reis, Marga (eds.) (2005): *Linguistic evidence. Empirical, theoretical and computational perspectives*. [Studies in Generative Grammar 85]. Berlin & New York: Mouton de Gruyter.

Kertész, András & Rákosi, Csilla (2008a): Introduction: The problem of data and evidence in theoretical linguistics. In: Kertész, András & Rákosi, Csilla (eds.): *New Approaches to Linguistic Evidence. Pilot Studies / Neue Ansätze zu Linguistischer Evidenz. Pilotstudien*. Frankfurt am Main: Lang, 9–19.

Kertész, András & Rákosi, Csilla (2008b): *Adatok és plauzibilis érvelés a nyelvészetben*. Debrecen: Kossuth Egyetemi Kiadó.

Kertész, András & Rákosi, Csilla (2008c): Megjegyzések a nyelvészeti adatok és evidencia problémájáról folyó vita jelenlegi állásához. I. rész. *Magyar Nyelv*, 274–286.

Kertész, András & Rákosi, Csilla (2008d): Megjegyzések a nyelvészeti adatok és evidencia problémájáról folyó vita jelenlegi állásához. II. rész. *Magyar Nyelv*, 385–402.

Kertész, András & Rákosi, Csilla (2012): *Data and evidence in linguistics: a plausible argumentation model*. Cambridge: Cambridge University Press.

Kertész, András & Rákosi, Csilla (2013): Az adattípusok integrációjának tudománymódszertani problémái az elméleti nyelvészetben. In: Kugler N., Laczkó K. & Tátrai Sz. (szerk.): *A megismerés és az értelmezés konstrukciói – nyelv, kultúra, irodalom.* Budapest. (megj. előtt)

Lehmann, Christian (2004): Data in linguistics. *The Linguistic Review* 21/3–4, 175–210.

Penke, Martina & Rosenbach, Anette (eds.) (2007): *What Counts as Evidence in Linguistics. The Case of Innateness.* [Benjamins Current Topics 7]. Amsterdam, Philadelphia: Benjamins.

Pérez Saldanya, Manuel (1996): Gramaticalització i reanàlisi: el cas del perfet perifràstic en català. In: Schönberger, Axel & Stegmann, Tilbert Dídac (ed.): *Actes del desè col.loqui internacional de llengua i literatura catalanes*. Volum 3. Barcelona: Publicacions de l'Abadia de Montserrat, 71–107.

Rescher, Nicholas (1976): *Plausible reasoning*. Assen & Amsterdam: Van Gorcum.

Segura-Llopes, Carles (2012): El passat perifràstic en català antic. Una revisió a partir d'estudi de corpus. *eHumanista/IVITRA* 2, 118–147.

http://www.ehumanista.ucsb.edu/eHumanista%20IVITRA/Volume%202/pdf/6%20Segura%20118-147.pdf. (utolsó letöltés: 2013.05.08.)

Soldevila, Ferran (1963–1968): L'ús del pretèrit perifràstic en la Crònica de Muntaner. In: *Estudis de Lingüística y de Filologia Catalanes dedicats a la memòria de Pompeu Fabra en el centenari de la seva naixença*. Volum 1. [Estudis Romànics 12]. Barcelona: Institut d'Estudis Catalans, 267–270.

Traugott, Elizabeth C. & Dasher, Richard B. (2004[2002]): *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Wasow, Thomas & Arnold, Jennifer (2005): Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496.

Nagy C. Katalin
MTA-DE Elméleti Nyelvészeti Kutatócsoport
4010 Debrecen
Pf. 47
nagykati@hist.u-szeged.hu