Online nganaszan történeti-etimológiai szótár

Szeverényi Sándor[1], Tóth Attila[2]

[1] Szegedi Tudományegyetem, Finnugor Nyelvtudományi Tanszék,
6722 Szeged, Egyetem u. 2.
szevers@hung.u-szeged.hu

[2] Szegedi Tudományegyetem, JGYPK Informatika Alkalmazásai Tanszék,
6725 Szeged, Boldogasszony sgt. 6.
attila@jgypk.u-szeged.hu

Kivonat: A bemutatóban a nganaszan nyelv online diakrón kognitív onomasziológiai szótár munkálatairól számolunk be. A szótár diakrón, mert a szókészlet történeti-etimológiai hátterét tárja fel, kognitív, mert az egymással összefüggő alakok közötti szemantikai kapcsolatokat is meghatározza, és onomasziológiai, mivel fogalmak felőli keresést, rendszerezést is lehetővé tesz. Mindezt úgy, hogy nem egy kész szótárat digitalizál, hanem olyan webes felületet hozunk létre, amely egyben a kutatás eszköze is.

1 A projekt célja

Projektünk újszerűsége egy történeti lexikográfiai probléma új típusú számítógépes feldolgozása. A szótár alapja már létezik, nyilvánossá a projekt végén, 2015 tavaszán fog válni. A projekt az OTKA támogatásával valósul meg.[1]

A munkálat nyelvészeti célja a nganaszan nyelv kognitív diakrón onomasziológiai szótárának kialakítása, a nganaszan szókincs rendszerezése szinkrón és diakrón szempontból (erről részletesebben [13]), olyan módon, hogy a későbbiekben a szótár más nyelvek adataival is ki tudjanak egészülni. Éppen emiatt a célkitűzések között szerepel a folyamatos javíthatóság és bővíthetőség biztosítása.A megvalósításhoz kapcsolódó technikai elvárások a következőkben foglalhatók össze: egy olyan szabad felhasználású, weben elérhető online felület, „eszköz” létrehozása, amely egyszerű módon jeleníti meg egy-egy lexéma formai, szemantikai tulajdonságait, történeti hátterét, valamint kapcsolatait más lexémákkal, és a megjelenített információk között összetett keresési kombinációkat tesz lehetővé.

2 A nganaszan nyelv

A nganaszan nyelv szókincsének és annak történetének dokumentáltsága tipikusnak mondható – a világ nyelveinek jelentős részéhez hasonlóan kevesen beszélik, hiányosan dokumentált és a beszélői kompetencia gyorsan tűnik el.

[1] A nganaszan nyelv diakrón kognitív onomasziológiai szótára (K100854).

Már az első lejegyzett nyelvi adatok is viszonylag késői időkből, a 18. század végéről származnak, s a módszeres nyelvi gyűjtés csak a 20. század utolsó évtizedeire vált általánossá. M. A. Castrén 19. századi gyűjtései ugyan történeti szempontból is jelentősek, ám a mennyisége nem teszi lehetővé, hogy külön történeti rétegként jelenítsük meg. Nganaszan írásbeliség nem alakult ki, mindössze egy gyakorlati szótár [10] és egy iskolás könyv [14] jelent meg. A kilencvenes években elsősorban Eugene Helimski, majd később tanítványa Valentin Guszev vezetésével történt szisztematikus nyelvi gyűjtés, melynek révén a nganaszan anyag mennyisége megsokszorozódott, ehhez magyar kutatók gyűjtései is hozzájárultak. Jelenleg a számunkra elérhető anyag mennyisége kb. 40-50 000 mondat.

Nem meglepő, hogy a nganaszan nyelv nagyon gyorsan halad az eltűnés felé. Beszélőinek száma a 2010-es oroszországi népszámlálási adatok szerint 125, a nyelvet anyanyelvi szinten beszélőké viszont ennek csak a töredéke lehet. Ez azt jelenti, hogy anyanyelvi kompetencia a projekthez nem áll rendelkezésre, jelentős mennyiségű, normalizált írásos korpusz pedig nincsen.

A nganaszan szókincs történeti háttere is csak részben feltérképezett, ez elsősorban a szókészlet szamojéd, uráli eredetű részére vonatkozik. Nincsen olyan korábban megjelent munka, amely a teljes nganaszan szókincs történetét, sajátosságait bemutatná, azaz a mára általánossá váló eljárás – egy nyelv vagy nyelvcsalád történeti-etimológiai szótárának digitalizálása, majd annak átdolgozása, frissítése, kiegészítése – a mi esetünkben nem lehetséges. Ugyanakkor annak sem látjuk értelmét, hogy napjainkban (csak) papíralapú szótárat készítsünk (noha az elmúlt időszakban a nemzetközi irodalomban van ilyenre példa, például [2, 3, 11]), illetve annak sem, hogy először elkészítsünk egy szótárat, s utána végezzük el a digitalizálást.

Mi megfordítottuk a sorrendet: előbb készítjük el a digitális verziót, s onnan lehet majd letölteni – a kívánt keresési eredményekkel – a nyomtatottat. Ehhez viszont olyan szerkezetet kellett kialakítani, amelyet lehetőség szerint a későbbiekben ne kelljen módosítani, csak finomítani, még akkor sem, amikor új nyelvek adatait dolgozzuk fel. Ennek megfelelően nemcsak az a feladat, hogy a nganaszan nyelvhez „passzoló” paraméterlistákat dolgozzunk ki, hanem a tipológiai szempontok is érvényesülni tudjanak.

2.2 A nganaszan korpusz

A nganaszan nyelvi anyagot zárt korpuszként kezeljük, ennek törzsanyagát az említett szótár adja (kb. 3500 címszó), illetve az azon alapuló angol változat [1]. Ezt az anyagot egészítjük ki olyan szócikkekkel, amelyek más forrásokban fordulnak elő. A történeti tárgyú munkák anyagát is külön-külön dolgozzuk fel, ezek legfontosabb forrásai: Janhunen 1976, Janhunen 1981, Helimskij 1997, [5, 6, 7]. Ezért gondoltuk, hogy célszerű lenne egy olyan szótár kialakítása, amelybe folyamatosan lehet „pakolni” az információkat, ha új közlések, publikációk jelennek meg, akkor azok anyagát rögtön be lehessen építeni az adatbázisba.

3 A szótár szerkezete

A szótár sajátos vonása, hogy a hangtörténeti jellemzők helyett a lexikológiai hátteret vizsgálja: definiálja a szóalakok közötti kapcsolatot, és a hozzájuk rendelhető jelentések közötti kapcsolatokat. Ennek megfelelően a szótárnak három fontos felülete van: a paraméterlisták („data”) felülete, a „form-concept” felület, és a „process-relation” felület.

3.1 A paraméterlisták (data)

A következő információcsoportok szerkeszthető rendszere található itt:

- nyelv / nyelvjárások: a rekonstruált (proto) nyelvek és az adatbázisban előforduló természetes nyelvek és nyelvjárások együttes listája;
- a szófaji rendszer: a jelentéssel együtt tárolt információ, jelenleg a nganaszan szófaji rendszerét tükrözi;
- irodalomlista: egyfelől az elsődleges adatokat tartalmazó munkákat, másfelől a szekunder hivatkozásokat tartalmazza;
- a szóalakok közötti kapcsolatok rendszere: a szóalkotási módok és azok alcsoportjai (összetétel, képzés, reduplikáció, kölcsönzés, folytonosság);
- opacitás: a motivációra vonatkozik, azaz átlátszó vagy átlátszatlan-e egy kifejezés;
- bizonyosság: a megállapított kapcsolat bizonyossága (biztos vs. bizonytalan);
- a szemantikai kapcsolatok rendszere: a rendszer nagyrészt a tübingeni kutatók által kidolgozott felosztást követi (például [4, 8], lásd lejjebb);
- jelentéscsoportok rendszere: a jelentéscsoportok rendszerét a Rapid Word Collection módszerét – amelyet kifejezetten dokumentációs nyelvészek számára dolgoztak ki a SIL munkatársai [12] – követve alakítottuk, illetve alakítjuk ki. Azért döntöttünk e felosztás mellett, mivel egyfelől az anyag szabadon felhasználható és adaptálható, másfelől a kategorizálás során hasonló kérdések merülnek fel, mint amikor terepmunkát végzünk, azaz egy gyakorlati szótári anyagot leginkább ez követ.
- speciális karakterek: egy újabb nyelv bekapcsolása azt is jelentheti, hogy új karakterre van szükség, itt könnyedén tudjuk előállítani a megfelelő karakterek, amelyek rögtön megjelennek a virtuális „billentyűzeten”.

Mindegyik csoport egyszerűen módosítható (bővíthető, ill. törölhető). Természetesen arra figyelemmel kell lenni, hogy például egy adott paraméter törlése (pl. nyelvjárás) milyen kapcsolatokban okoz változást (pl. az adott nyelvjárásba tartozó lexémák).

3.2 Szóalakok és jelentések (form & concept)

Ez a rész szolgál a szóalakok és jelentések bevitelére, katalogizálására és a lexéma-jelentés kapcsolatok létrehozására. Ez azt jelenti, hogy egy szóalakot csak egyszer tárolunk el, homonímia esetén sem szükséges az alakot újra rögzíteni. A jelentéseknél

hasonló a helyzet, azzal a különbséggel, hogy a jelentéseket minden esetben úgy kell megadnunk, ahogyan a forrásban szerepelnek, így például a ’mountain’ jelentés háromszor szerepel jelenleg a szótárban:

’mountain ridge, mountain range’
’mountain, rock’
’mountain, hill, ridge’

Egy ’mountain’ részleges egyezéses keresés kiadja mindhárom találatot, s ha teljesen biztosak akarunk lenni abban, hogy minden találat megjelent-e, akkor a ’mountain’ jelentéscsoportját (jelenleg LAND) is lehet használni.

3.3 Lexémák és szemantikai kapcsolatok

Saját szerkesztői felülete van az egyes lexéma+jelentés párok közötti alaki és szemantikai kapcsolatoknak (process – relation), ugyanitt lehet a változás irányát is meghatározni (source – target). Ez felveti azt a kérdést, hogy a jelentésváltozás és a szinonímia között megállapítható-e a határ.

A szóalkotási eljárások (process) jelenleg a nganaszan szóalkotási módokat tartalmazza (képzés, átvétel, összetétel, lexikai folytonosság stb.), illetve ezek alcsoportjait. A jelentések közötti kapcsolatokat két nagy csoportja a metaforikus (hasonlóságon alapuló), illetve a metonimikus (kontiguitáson alapuló) kapcsolatok.

Természetesen egy kapcsolatot több minősítéssel is el lehet látni. Amit pedig a minősítésekkel nem lehet megadni, azt a „comment” részben lehet megmagyarázni. Fontos, hogy a rendszer a formai és a jelentésbeli változásokat, kapcsolatokat együtt láttatja, a diakrón kognitív onomasziológiai munkálatoknak ez az egyik alapvető célja.

Mivel a kapcsolatok meghatározása gyakran nem egyértelmű, vagy csak nagyon „leegyszerűsítve” adja vissza a tényleges relációkat, ezért a „comment” résznél lehetőség van szöveges kiegészítésre.

Ezáltal gyakorlatilag szóláncokat tudunk létrehozni, be tudjuk mutatni egy adott szótő eredetét, más nyelvekben való megjelenését, származékait, jelentéseit, s azok viszonyait.

3.4. Keresés

Az elmondottakat az *ə̂ntə̂j* ’boat’ > *ŋənduj* ‘a kind of boat’ > *tuu ŋənduj* ’steamboat, steamer, steamship’ szólánccal szemléltetjük. A nganaszan *ŋənduj* ‘a kind of boat’ szóra keresünk rá. Elsődleges forrása az említett Kosterkina et al. (2003) szótár [10]. A jelentést besoroltuk a TRAVEL és a FISHING kategóriákba. Ha rákeresünk a *ŋənduj* szó, akkor a következő lényeges információkat kapjuk:

- a *ŋənduj* forrása a proto-szamojéd rekonstruált *ə̂ntə̂j* ’boat’. Ennek forrása Janhunen etimológiai szótára;
- a *ŋənduj* és a *ə̂ntə̂j* szóalakok között kapcsolat lexikai folytonosság (azaz a nganaszanban egy korábbi nyelvállapotra rekonstruálható alak a hangváltozásokat leszámítva változatlanul meg);

- a *ŋənduj* és a *ə̂ntə̂j* szóalakok közötti kapcsolat leginkább a konceptuális/fogalmi azonosság kategóriájába tartozik, mivel mindkettő csónakot jelent;
- a *ŋənduj* 'boat' szóalak + jelentés kapcsolat részleges forrása újabb elemeknek, így például a *tuu ŋənduj* ' steamboat, steamer, steamship' szókapcsolatnak;
- a *tuu ŋənduj* 'steamboat, steamer, steamship' szóalak + jelentés forrásai között megjelenik a *tuj* 'fire' szó is. A *tuu ŋənduj* összetételt szóalkotási szempontból összetételnek minősítjük. A *tuu* a *tuj* szóalak genitívuszi alakja (ezt az információt a comment részben tudjuk tárolni). Természetesen a *tuu ŋənduj* forrásai között a *tuj* is megjelenik;
- A *ŋənduj* 'a kind of boat' és a *tuu ŋənduj* 'steamboat, steamer, steamship' közötti szemantikai kapcsolat egyfajta fogalmi hasonlóságon alapuló specializáció, a csónak jármű egy speciális fajtájára utal, ezért a metaforikus kapcsolatok közül a fogalmi hasonlóság mellett a taxonomikus alárendelés is szerepel a minősítések között.

4 A technikai háttér

Mivel a cél olyan online rendszer kifejlesztése volt, amely adattartalma folyamatosan fejleszthető és felhasználása minél szélesebb kör számára elérhető, így a webes alkalmazás a legkézenfekvőbb megoldás. Ezáltal a felhasználói és az adminisztrátori funkciók elvégzéséhez is elég egy böngésző. Ez jelentősen megkönnyíti a bővítési, további nyelvekkel való kiegészítési munkafolyamatot.

Alapvető elvárás a rendszerrel szemben, hogy az adattartalom dinamikusan változtatható, bővíthető legyen úgy, hogy az adatok redundanciáját elkerüljük. Így a rendszer alapját egy olyan SQL adatbázis képezi, amely központi magját a szóalak és jelentés párok alkotják, illetve az ezekből képezett formális és szemantikai kapcsolatok. Azaz külön egységként tároljuk a szóalakokat és a jelentéseket, az ezek közötti kapcsolatot, valamint az így képzett párok közötti átmeneteket. Ez a modell alkalmas arra, hogy bizonyos szóalakok (illetve jelentések) több jelentéssel (illetve szóalakkal) is párt alkossanak, így a poliszém és a homonim alakok redundanciamentesen jól ábrázolhatók. Továbbá az ezeket jellemző attribútumok lehetséges értékei szintén külön tároltak, így ezek bővítése könnyen elvégezhető.

Egy ilyen rendszerben elemi elvárás, hogy az alkalmazás képes legyen a tartalmazott nyelvek speciális karaktereinek a kezelésére, illetve olyan felhasználói felületet nyújtani, ahol az ilyen karakterek könnyen beilleszthetőek. Mivel a szerzők célja a rendszert további nyelvekre is kibővíteni, így ennek kezelését rugalmasan kell megoldani. Emiatt egyrészt az adattárolás UTF-8 kódolással történik, valamint az adatbázisban külön tárolásra kerülnek a speciális karakterek és azok kódjai is. Másrészt a speciális karakterek bevitelét a felhasználói felületen egy virtuális billentyűzet segíti, amelyen szereplő karakterek dinamikusan állnak össze az adatbázis ilyen karaktereit tartalmazó tábla tartalma alapján.

5 Tervek

Szótárunkkal azokhoz a kutatásokhoz kívánunk a jövőben kapcsolódni, amely leginkább a lexikális tipológia, s annak különösen a diakrón ágához tartozik. Koch és Marzo [9] szerint a lexikalizáció formai és kognitív motivációjának diakrón tipológiai rendszerezése a következők miatt fontos:

(i) lehetővé teszi az egyes nyelvek motivációs „profiljának” megalkotását;
(ii) lehetővé teszi nyelveken átívelő tendenciák és idioszinkráziák megállapítását (Vannak-e „transzparensebb” vagy kevésbé transzparens nyelvek? Vannak-e „metaforikusabb” nyelvek?);
(iii) lehetővé teszi nyelveken átívelő és nyelvspecifikus motivációs preferenciák megállapítását.

Ezért célunk, hogy az adatbázis további nyelvekkel, s adatokkal bővüljön, s a munka a projekt lejárta után is folytatódjon.

Hivatkozások

1. Bradley, J., Wagner-Nagy, B.: Nganasan–English Dictionary. Ms. Wien: Hamburg. (2013)
2. Fortescue, M., Jacobson, S, Kaplan, L.: Comparative Eskimo Dictionary . Alaska Native Language Press, Fairbanks (1994, 20122)
3. Fortescue, M.: Comparative Chukotko-Kamchatkan Dictionary. Trends in Linguistics. Documentation. Mouton de Gruyter, Berlin: New York (2005)
4. Gévaudan, P.: Typologie des lexikalischen Wandels. Stauffenburg, Tübingen. (2007)
5. Helimski, E.: Die matorische Sprache. SUA 41. JATE, Szeged (1997)
6. Janhunen, J.: Samojedischer Wortschatz. Castrenianumin toimitteita 17, Helsinki (1977)
7. Janhunen, J.: Uralilaisen kantakielen sanastosta. JSFOu 77. (1981) 219–274
8. Koch, P.: Lexical typology from a cognitive and linguistic point of view. In Haspelmath, Martin, König, Ekkehard, Oesterreicher, Wulf, Raible, Wolfgang (Hrsg.): Linguistic Typology and Language Universals = Handbook of Linguistics and Communication Science 20/2. Mouton de Gruyter, Berlin. (2001) 1142–1176
9. Koch, P., Marzo, D.: A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. Studies in Language 31:2 (2007) 259–291.
10. Kosterkina, N. T., Momde, A. Č., Ždanova, T. Ju. [Костеркина, Н. Т., Момде, А. Ч., Жданова, Т. Ю.]: Словарь нганасанского-русский и русско-нганасанский, Филиал издательство «Просвещение», Санкт-Петербург (2001)
11. Nikolaeva, I.: A Historical Dictionary of Yukaghir. Trends in Linguistics. Documentation. Mouton de Gruyter, Berlin: New York (2006)
12. Rapid Word Collection http://www.rapidwords.net/ (2013. november 28.)
13. Szeverényi S.: Mire jó egy nganaszan online diakrón kognitív onomasziológiai szótár? Nyelvtudományi Közlemények 108 (2012) 197–218
14. Žovnickaya, S. N. [Жовницкая, С. Н.]: Букварь, Санкт-Петербург, Просвещение (2001)