Nyelvtanulást elősegítő feladatok automatikus előállítása finn és magyar nyelvekre

Ferenczi Zsanett

Pázmány Péter Katolikus Egyetem
ferenczi.zsanett@hallgato.ppke.hu

Kivonat Nyelvtanulók számára gyakran nehézséget okoz agglutináló jegyekkel rendelkező nyelvek esetén az egyes lemmák különböző esetragokkal ellátott alakjainak előállítása, valamint annak elsajátítása, hogy egy adott környezetben pontosan mely esetragot szükséges használni. Ezen kívül a nyelvek egyedi jellegzetességei is kisebb-nagyobb akadályokat gördítenek a nyelvtanulók elé. A nyelvtanulás elősegítése érdekében meghatároztunk néhány feladattípust, majd automatikus módszereket használva és bizonyos nyelvi mintázatokat kiaknázva példamondatok ezreiből egy nyelvtanuló alkalmazást hoztunk létre finn és magyar nyelvet tanulók számára. Ezen folyamatot jelen cikkben mutatjuk be.
Kulcsszavak: számítógéppel támogatott nyelvtanulás, nyelvtanulás, szókincsfejlesztés, digitális szókártyák, finnugor nyelvek

1. Bevezetés

Az idegen nyelvek tanulásának hátterében több ok is állhat. Ha valaki külföldön szeretne elhelyezkedni, szinte elengedhetetlen az adott ország nyelvének ismerete, a diploma megszerzéséhez ma már szükség van egy államilag elismert nyelvvizsga megszerzésére, valamint pusztán az idegen nyelvek iránti érdeklődés is elengedő ok lehet arra, hogy egy nyelvet elsajátítsunk.

A digitális eszközök az oktatás és nyelvoktatás területén is egyre gyakrabban vannak jelen. A kontakt oktatási környezetet már sok esetben felváltja az online keretek között zajló oktatás. Az elmúlt években a járványügyi helyzet miatt sok nyelviskola is arra kényszerült, hogy teljesen online, esetleg hibrid (azaz félig online, félig tantermi) tanfolyamokat indítson.

Egy hagyományos nyelvkurzuson, esetleg iskolai nyelvoktatás során a nyelvoktató feladata az, hogy a diákokat elegendő gyakorlattal, nyelvi feladattal lássa el, amelynek segítségével a tanulók a nyelvórán újonnan elsajátított tananyagokat önállóan is begyakorolhatják. Elegendő mennyiségű gyakorlatot manuálisan előállítani időigényes feladat lenne, ezt az oktatók sokszor úgy próbálják meg elkerülni, hogy több tankönyvből, munkafüzetből, esetenként digitális erőforrásokból dolgoznak a nyelvórák során. Ezen túlmenően azonban a feladatokra adott válaszokat is nagy munka egyetlen tanárnak kézzel ellenőrizni, amennyiben az adott gyakorlófeladatokhoz nem tartozik olyan megoldókulcs, melyet a tanulók használhatnának a saját maguk által megírt feladatok ellenőrzéséhez.

Az uráli nyelvcsaládhoz tartozó finn és magyar nyelv tipológiailag agglutináló nyelvnek számítanak, gazdag morfológiájuk miatt viszonylag nehézkes az elsajátításuk. A finn azonban nem nevezhető prototipikusan agglutináló nyelvnek. Ez olyan morfofonológiai változásoknak (pl. fokváltakozás) köszönhető, amelyek a nyelvtanulók dolgát még inkább megnehezítik, és még több gyakorlásra adnak okot (Simon, 2015). Ezen nyelvek tanulásakor nagy erőfeszítést igényel a magánhangzó-harmónia és bizonyos morfofonológiai változások megértése, helyes használata. Nem csak a paradigmák között fennálló változatosság, eltérés jelent problémát tanuláskor, hanem az egyes szavak paradigmáinak mérete, az esetrendszer kiterjedtsége is. A magyar esetek száma (irodalomtól függően) 17 és 28 közé tehető (vö. Tompa (1961), Antal (1961), Kiefer (1987, 2006, 2018)), míg a finnben általában 14 vagy 15 nyelvtani esetet szokás megemlíteni (Hakulinen és mtsai, 2004). Ezek teljes elsajátítása tehát jóval több időt vesz igénybe ezen agglutináló nyelvek esetében, mint például az indoeurópai nyelvek közé tartozó angol 2-3 esete, a német 4 esete vagy az orosz 6 esete.

Máté (1999) felmérése alapján a magyar nyelvet tanuló finn anyanyelvűek számára nehézséget jelent még a határozott és határozatlan igeragozás elsajátítása, az igekötők, valamint a magyar birtokos szerkezetek helyes használata is.

Ugyanakkor a finnül tanulók is különböző nehézségekbe ütköznek nyelvtanuláskor. Karlsson és Chesterman (2008) szerint a legtöbb tanuló számára megdöbbentő, hogy a finn szókincs szinte semmilyen másik nyelv szókincséhez nem hasonlít, és nagyon bonyolult szabályok határozzák meg, hogy a mondat tárgya a lehetséges esetek közül éppen melyikben áll egy adott mondatban. Az is újdonságot jelent számukra, hogy a finn három múlt időt használ a magyar sztenderdben használt egy múlt idővel szemben.

2. Kapcsolódó irodalom

Ezen kihívást jelentő feladatok egyszerűbbé tételére, automatizálására már különféle megoldásokat kínáltak különböző nyelvek esetén. A számítógéppel támogatott nyelvtanulás (angolul *Computer-assisted language learning*, röviden CALL) célja, hogy fellendítse azon digitális eszközök, alkalmazások kiépítését, amelyek segítségével a tanulók a nyelvtudásukat, nyelvi készségeiket fejleszteni tudják. A CALL egyik tipikus megnyilvánulási formája az úgynevezett *fill in the blank*, azaz behelyettesítéses feladatok, amelyeket különböző nyelvtechnológiai eszközök segítségével akár automatikusan is létrehozhatunk és kiértékelhetünk. Ilyen feladatra láthatunk példát az 1. ábrán.

The priest was innocent ____ he could not prove it.

1. ábra: Behelyettesítéses feladat.

A tanulók önállóan gyakorolhatnak számos idegen nyelvet, illetve főleg azok szókincsét például olyan platformokon, mint a Memrise[1], Busuu[2] vagy a Duolingo[3]. Ezen applikációknak csak egy része ingyenes, általában az összes általuk kínált funkció csak előfizetéssel vehető igénybe. Ezen kívül a fenti platformok csak a legtöbb beszélővel rendelkező nyelveket ölelik fel, mint például az angol, francia, orosz és kínai. A három applikáció közül csak a Duolingo tartalmaz finnugor nyelveket, a finnt és a magyart.

Vannak kifejezetten a kisebb uráli nyelveket és azok tanulását fellendíteni kívánó kutatások, melyek automatikus módszerekkel hoznak létre ingyenes tananyagokat a tanulók számára (Uibo és mtsai (2015), Antonsen és mtsai (2009)). A Revita projekt (Katinskaia és mtsai, 2017) keretein belül olyan alkalmazást fejlesztenek, amely segítségével szintén néhány kisebb finnugor nyelvet vagy akár a finnt is gyakorolhatjuk sok nagyobb nyelv (pl. orosz, svéd) mellett. A Revita online felületén a legalább középfokú nyelvtudással rendelkező felhasználó behelyettesítéses feladatokat oldhat meg. Az oldalon megjelenik egy egybekezdésnyi célnyelvi szöveg, amelyből bizonyos szavakat automatikus módszerekkel eltávolítottak. A szöveget a nyelvtanulónak kell rekonstruálnia úgy, hogy beírja a megadott lemma alapján a mondatból hiányzó alakot, vagy kiválasztja a legördülő menüben feltűnő elemek közül a szövegbe leginkább illeszkedő szót. Egyes elemek esetén a tanuló hallás utáni szövegértését mérik: ez esetben egy kis hangfájlt kell meghallgatni, majd a hallott szót megadni a szövegbeviteli mezőben. Ezen feladatok esetén azonban teljesen tetszőleges módon távolítják el a szavakat a szövegből. Egy nyelv tanulásakor a tanuló egyre több nyelvtani szabállyal bővíti tudását, és ezen nyelvtani egységek elsajátítását az arra kiélezett feladatokkal tudja legkönnyebben begyakorolni. Fontos, hogy a Revita felületét használók már valamilyen szinten ismerjék, beszéljék az adott nyelvet, mivel ezen projektben nem osztják fel nyelvtani típusfeladatokra a szövegeket, és nem a tanuló nyelvtudásának aktuális szintje határozza meg a feladatok nehézségét.

Jelen cikkben bemutatjuk azt az online nyelvtanuló alkalmazást, amely segítségével a finn, illetve magyar nyelvet tanulók elmélyíthetik tudásukat, bővíthetik szókincsüket, és begyakorolhatják az ezen nyelvek esetén leggyakrabban nehézséget okozó nyelvtani szabályosságokat.

3. Számítógéppel támogatott nyelvtanulás

3.1. Szókincs

Különböző stratégiák léteznek az idegen nyelvű szókincs minél hatékonyabb bővítésére. Ezek egyike a papíralapú vagy virtuális szókártyák használata.

A szókártyák célja a nyelvtanuló szókincsének bővítése, új szavak, kifejezések elsajátítása. Ilyen kártyákat több módon is létrehozhatunk: hagyományosan a célnyelvi (L2) szót, kifejezést felírjuk egy kártya egyik oldalára, míg a hátoldalra

[1] `https://www.memrise.com/`
[2] `https://www.busuu.com/`
[3] `https://www.duolingo.com/`

vagy a forrásnyelvi (L1) vagy a célnyelvi (L2) definíció kerül. Elgort (2013) kutatása azt mutatta ki, hogy kezdő nyelvtanulók szignifikánsan jobb eredményt érnek el a forrásnyelven való megfelelők segítségével, míg a haladóbb tanulók esetében a különbség L1, valamint L2 definíciók használata között nem ennyire látványos. Jo (2018) felmérése is azt mutatja, hogy L1 definíció segítségével magasabb pontszámot értek el a tanulók, mint L2 definíciókkal.

Több kutatás is létezik, amely a papíralapú szókártyákat veti össze a virtuális szókártyák használatával, és ezek közül például Kilickaya és Krajka (2010), valamint Basoglu és Akdemir (2010) is kimutatta, hogy a virtuális szókártyákat használó csoportok teljesítménye felülmúlta a papíralapú kártyákat használókét.

Ezen kutatások alapján alkalmazásunkban virtuális szókártyákat hoztunk létre mind kezdő, mind haladóbb szinteken lévő tanulók számára. A szókártyák használata előtt a tanuló döntheti el, hogy a kártyák hátoldalán a kérdéses elem fordítását vagy annak célnyelvi definícióját szeretné-e látni. Az új szóanyag elsajátítása utáni tesztfázis a produktív előhívási készségeket (Laufer és mtsai, 2004) fejleszti. Ez azt jelenti, hogy a tanulónak emlékeznie kell a korábban látott célnyelvi szóra, és meg kell azt adnia a megfelelő szövegbeviteli mezőben, mindezt úgy, hogy a kérdéses szónak vagy a forrásnyelvi megfelelőjét vagy a célnyelvi definícióját jelenítjük meg számára.

3.2. Nyelvtan

A különböző nyelvtani szerkezeteket a tanuló a korábban is bemutatott behelyettesítéses feladatok segítségével gyakorolhatja, melyet finn és magyar számítógépes nyelvfeldolgozó eszközökkel megtámogatva állítottunk elő.

Alkalmazásunk létrehozásakor a nyelvtan azon részeire fókuszáltunk, amelyek a legnagyobb nehézséget okozzák magyarul tanuló finn, illetve finnül tanuló magyar anyanyelvűek számára Máté (1999) és Karlsson és Chesterman (2008) megfigyelései alapján (lásd 1. fejezet). Összesen három finn és három magyar nyelvtani egységre dolgoztunk ki olyan mintázatokat, melyek segítségével a tanulók példák ezrein keresztül tudják gyakorolni a különféle nyelvi jelenségeket. A finn esetében az egyik feladattípus a tárgy esetének kiválasztását gyakoroltatja, egy másik a három múlt idő közötti különbségtételt segíti elő, míg a harmadik a finn passzívum képzését fejleszti. A magyar nyelvtant érintő feladattípusok az igekötők gyakorlása, a határozott és határozatlan igeragozás közötti különbségek megértése, valamint a birtokos szerkezetek elsajátítása köré szerveződnek.

Egy feladatsor összeállításakor az első lépés az, hogy a korábban összegyűjtött példamondatok közül kiszűrjük a feladat típusának megfelelő mondatokat, amelyben a vizsgált nyelvi jelenség megjelenik. Ezután a kérdéses szót, szavakat eltávolítjuk a mondatból. Az eltávolított szónak bizonyos feladatok esetén megjelenítjük a lemmáját, hogy a tanulónak csak a megfelelő esetben, időben, személyben és/vagy számban kelljen elragoznia a szót. Az egyik feladattípusban (egészen pontosan a magyar igekötők esetén) a tanulónak egy zárt halmazból kell kiválasztania a megfelelő elemet, ilyenkor értelemszerűen nem adjuk meg a hiányzó szó lemmáját.

A szövegbeviteli mezők kitöltését és az űrlap elküldését követően a kiértékelés automatikusan történik, a rendszer összeveti a felhasználó válaszát a mondatban eredetileg szereplő kifejezéssel, és megadja a helyes válaszokat ott, ahol azok nem egyeztek meg.

4. Finnugor nyelvtanuló applikáció

4.1. A felhasznált adatok

A virtuális szókártyák és a nyelvtani gyakorlatok alapjául egy MySQL adatbázis szolgál. A finn és magyar szavakat, többszavas kifejezéseket és mondatokat tartalmazó tábla neve `Entity`. Ebben automatikus módszerekkel kigyűjtött finn és magyar lemmák, definíciók, példamondatok szerepelnek olyan erőforrásokból, mint a magyar[4], finn[5] és angol[6] Wiktionary, a magyar (Miháltz és mtsai, 2008) és finn WordNet (Lindén és Carlson, 2010) és az OPUS korpusz (Tiedemann és Nygaard, 2004).

A fent említett erőforrásokat különböző nyelvtechnológiai eszközök és eljárások segítségével használtuk fel a kétnyelvű szópárok kinyeréséhez, valamint az erőforrásokban megtalálható szinonimák, példamondatok és definíciók kigyűjtéséhez.

A Wiktionary különböző verzióiból az Ács és mtsai (2013) által létrehozott `wikt2dict` eszközt használtuk fel, amellyel kétféle módon juthatunk kétnyelvű szólistákhoz. Egyrészt a Wiktionary szócikkekben megtalálható fordítási táblákból gyűjt ki fordításokat azon szócikkek esetén, amelyeknél a címszó nyelve megegyezik a Wiktionary nyelvével, másrészt egy harmadik nyelvet felhasználva ugyanezen táblák segítségével úgynevezett háromszögelési módszerrel újabb szópárokkal bővíti a szólistákat. Ez a módszer azon az elképzelésen alapul, hogy ha egy nyelv egy bizonyos szavát lefordítjuk két másik nyelvre, akkor vélhetően ezek a fordítások egymás fordításainak is tekinthetők. Harmadik nyelvként jelen esetben az angolt használtuk, mivel ez a Wiktionary rendelkezik a legtöbb szócikkel.

A Wiktionaryben nem csak a fordítási táblákban találhatunk finn–magyar fordításokat, hanem akkor is, amikor a finn Wiktionaryben magyar, illetve a magyar Wiktionaryben finn szavakra keresünk rá. Ezen párokat a `wikt2dict` eszköz nem gyűjti össze, így saját algoritmust írtunk ennek kiaknázására. Megoldásunk a finn és magyar nyelvű Wiktionaryk dumpjait járja be, és a kétnyelvű szólisták mellett a Wiktionary nyelvével megegyező nyelvű szócikkekből kinyert példamondatokat és definíciókat is eltárolja. A saját fejlesztésű eszköz kódja Creative Commons Attribution-ShareAlike 4.0 licenc alatt szabadon elérhető[7].

A finn és magyar WordNetekben szereplő synsetek (szinonimahalmazok) azonosítója lehetővé teszi, hogy kapcsolatot teremtsünk ezen egynyelvű erőforrások

[4] `https://hu.wiktionary.org`
[5] `https://fi.wiktionary.org`
[6] `https://en.wiktionary.org`
[7] `https://github.com/ferenczizsani/wiktionary_parser`

elemei között. Az egymásnak megfeleltetett synseteket feloldva és az egyes lemmákat összekapcsolva szópárok ezreihez jutunk. Ezen túl a magyar WordNetben jelen lévő példamondatokat is eltárolja algoritmusunk, melynek kódja Creative Commons Attribution-ShareAlike 4.0 licenc alatt szabadon hozzáférhető[8].

Az OPUS-ban található finn és magyar, szavak szintjén párhuzamosított szótárak számos fordítási párt tartalmaznak. Ezen listák elemeit szükséges volt lemmatizálásnak alávetni, mivel ez az erőforrás sok esetben különböző esetragokkal ellátott, de lemmájukat tekintve ismétlődő szópárokat tartalmazott. A finn és magyar szópárhuzamosításokat kigyűjtő algoritmus kódja szabadon elérhető CC BY-SA 4.0 licenc alatt[9].

Az `Entity` tábla tartalmazza a fent bemutatott módszerekkel kinyert entitásokhoz az entitás nyelvét, szófaját és annak típusát (amely lehet lemma, többszavas kifejezés vagy akár mondat is). A tábla felépítését az 1. táblázat foglalja össze (`id` = azonosító, `text` = szöveg, `lang` = nyelv, `upos` = szófaji címke, `type` = típus).

id	text	lang	upos	type
145	gyermek	2	NOUN	lemma
436	A szülők közvetlen leszármazottja.	2	NONE	sentence
733	A bál tánccal zárult.	2	NONE	sentence
918	Tiikeri on iso kissa.	1	NONE	sentence

1. táblázat. Az `Entity` tábla felépítése.

Az entitások között bizonyos kapcsolatok jöhetnek létre, amelyeket a `Relation` tábla tárol. Ez rögzíti a reláció típusát (pl. célnyelvi megfelelője, definíciója, példamondata, szinonimája) és a relációban részt vevő két entitás azonosítóját.

A nyelvtani gyakorlatok automatikus létrehozásához a különböző erőforrások magyar és finn példamondatait használtuk fel. Az adatbázisban szereplő finn nyelvű példamondatok száma 29.087, a magyar példamondatok száma 20.158. Ezen mondatok összegyűjtése után az alábbi kritériumok mindegyikének megfelelő adatokat további elemzésnek vetettük alá:

- legalább három szóból álló mondatok;
- nagybetűvel kezdődő mondatok;
- mondatvégi írásjelet (`.`, `!` és `?`) csak a mondat végén tartalmazó mondatok;
- matematikai jeleket (`+`, `=`) nem tartalmazó mondatok.

Az így megmaradt egyedi mondatok száma a finn nyelv esetén 18.043, míg a magyar mondatok száma 17.450. Ezen mondatok ezt követően tokenizáláson, lemmatizáláson, valamint morfológiai és függőségi elemzésen estek át.

[8] `https://github.com/ferenczizsani/connect_wordnets`
[9] `https://github.com/ferenczizsani/opus_extractor`

A magyar nyelvre az `emtsv` eszközt (Indig és mtsai, 2019), míg finn nyelvre az `omorfi` (Pirinen, 2015) és az `uralicNLP` (Hämäläinen, 2019) eszközt használtuk fel. Ezek kimenete egységesen a Universal Dependencies CoNLL-U[10] formátumát követi, így az adatok átjárhatósága adott.

A kinyert adatokat szintén a fenti adatbázisban tároljuk el. A `TokenAnalysis` tábla az egyes tokeneket és azok elemzését (lemma, `lang` = nyelv, `upos` = szófaj, `feats` = egyéb morfológiai, morfoszintaktikai jegyek) foglalja magában (lásd 2. táblázat). Az `Analysis2Sentence` tábla a mondatok és a bennük előforduló tokenek közötti kapcsolatokat tárolja, együtt a token mondatban elfoglalt helyével (= `token_position`), valamint a dependenciaelemzés kimenetével, azaz a szintaktikai címkével (= `deprel`) és a szülőcsomópont pozíciójának számával (= `head`), lásd 3. táblázat.

token_id	token	lemma	lang	upos	feats
13	.	.	2	PUNCT	_
56	kissa	kissa	1	NOUN	Case=Nom|Number=Sing
102	A	a	2	DET	Definite=Def|PronType=Art
479	bál	bál	2	NOUN	Case=Nom|Number=Sing
480	tánccal	tánc	2	NOUN	Case=Ins|Number=Sing
481	zárult	zárul	2	VERB	Definite=Ind|Mood=Ind| Number=Sing|Person=3| Tense=Past|VerbForm=Fin| Voice=Act

2. táblázat. A TokenAnalysis tábla felépítése.

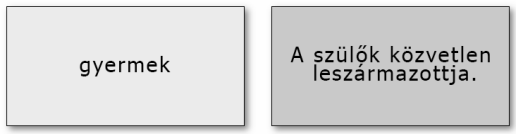
sentence_id	token_id	token_position	deprel	head
733	102	1	DET	2
733	479	2	SUBJ	4
733	480	3	OBL	4
733	481	4	ROOT	0
733	13	5	PUNCT	0

3. táblázat. Az Analysis2Sentence tábla felépítése.

[10] `https://universaldependencies.org/format.html`

4.2. Szókártyák

Mivel az adatbázisban a fordítási párok közötti, valamint a lemmák és azok definíciói közötti relációkat is eltároltuk, ezek felhasználhatók egy- vagy kétnyelvű virtuális szókártyák létrehozásához. Az alkalmazásunkban létrehozott virtuális szókártyák egyik oldalán tehát a célnyelvi szó, lemma található, míg a másik oldalon a tanuló kiválaszthatja, hogy a szó definíciója a célnyelven vagy a forrásnyelven szerepeljen-e (lásd 2. és 3 ábra).



2. ábra: Egy virtuális magyar nyelvű szókártya célnyelvi definícióval.

3. ábra: Egy virtuális magyar nyelvű szókártya forrásnyelvi definícióval.

A szókártyákkal való tanulás folyamata két részre osztható. A gyakorló fázisban a tanuló először begyakorolhatja, áttekintheti a számára eddig ismeretlen szavakat. Ebben a fázisban a tanuló a virtuális szókártya mindkét oldalát megtekintheti, és amikor úgy érzi, sikerült megjegyeznie az adott elemet, továbbléphet a következő kártyára. A tesztfázisban csak a kártyák definíció oldala jelenik meg véletlenszerű sorrendben, és a nyelvtanulónak meg kell adnia a definíció által leírt fogalmat.

Ezen kártyákat automatikusan generáljuk az adatbázis megfelelő tábláit felhasználva, és a válaszok is automatikusan kerülnek kiértékelésre. Fontos megjegyezni, hogy egy választ akkor fogad el helyesnek a rendszer, ha az teljes egészében megegyezik az első fázisban elsajátított fogalommal. Szinonimák megadására még nincs lehetőség, ezt a folyamatot a projekt egy későbbi szakaszában tervezzük megvalósítani.

4.3. Nyelvtani gyakorlatok

A finn és magyar nyelvtan három-három aspektusára hoztunk létre nyelvtani gyakorlófeladatokat, melyet ebben a fejezetben mutatunk be részletesen.

Finn feladatok A finn mondat tárgya négyféle esetben állhat a morfológiai és függőségi elemző kimenetét megfigyelve: nominatívuszban, partitívuszban, akkuzatívuszban és genitívuszban. Komplex szabályok határozzák meg, hogy ezen négy eset közül éppen melyiket veszi fel a tárgy. Az adatbázisban szereplő mondatok közül bizonyos nyelvi mintázatokat felhasználva automatikusan kiválaszthatjuk azokat, amelyek tartalmaznak legalább egy tárgyat, amely a fenti esetek egyikében áll. Az erre vonatkozó szabályt a következőképp fogalmazhatjuk meg: válasszuk ki azon finn nyelvű mondatokat, amelyekben található olyan névszó, amely a függőségek tekintetében a `DOBJ` jegyet kapta, és morfológiai jegyei között megtalálható az alábbiak egyike: `Case=Nom`, `Case=Par`, `Case=Acc` vagy `Case=Gen`. Összesen 7.088 mondat felel meg ennek a kitételnek, és ezen mondatokból a megfelelő tokent eltávolítva, azt egy szövegbeviteli mezővel helyettesítve, valamint a hiányzó token lemmáját és számát (egyes vagy többes szám) megadva előállítható a finn tárgyra vonatkozó feladatok gyűjteménye.

A finn nyelvben háromféle múlt időt különböztethetünk meg: imperfektumot, perfektumot és pluszkvamperfektumot. Az igék múlt idejű alakjainak előállítását a példamondatokból eltávolított igék megfelelő alakjának behelyettesítésével gyakorolhatja a tanuló. Az összetett múlt idők (a perfektum és a pluszkvamperfektum) az *olla* segédigéből és egy múlt idejű particípiumból tevődnek össze. Ehhez a feladathoz olyan példákat gyűjtünk ki az adatbázisból, amelyekben vagy imperfektumban áll az ige, vagy amelyek tartalmazzák az *olla* segédigét és egy múlt idejű melléknévi igenevet.

Ezen kitétel után megfigyeltük, hogy a segédige és a particípium nem feltétlenül állnak egymás mellett, így az imperfektum alakot azonnal ki lehet zárni a lehetséges múlt idők közül, mivel két törölt szó is megjelenik a mondatban (lásd 4. ábra). A feladat célja éppen az, hogy a tanuló a három igeidő közül kiválassza a legmegfelelőbbet az adott kontextusban, ezért ezen mondatokat jelen esetben ki kell zárnunk a példák közül. Ezt pillanatnyilag egy olyan további feltétel bevezetésével oldottuk meg, amely szerint a particípiumnak közvetlenül az *olla* ragozott alakja után kell következnie.

4. ábra: Példafeladat összetett múlt idő esetén.

A fennmaradó példák száma a szűrést követően 5.133. Ezekből az igealakokat (és esetenként a particípiumot) eltávolítottuk, és a szövegbeviteli mező után zárójelben feltüntettük az ige első infinitívuszi alakját. Mivel — a magyarhoz hasonlóan – a finn is pro-drop nyelv, a helyes alak előállításához szükség van a mondat alanyának számára és személyére vonatkozó információkra is.

A passzívum fontos szerepet tölt be a finn nyelvben. Ezt mutatja az is, hogy az adatbázisban minden tizennegyedik finn példamondat passzív alakú igét tartalmaz. Ez a jelenség a beszélt nyelvben még gyakoribb, ugyanis a többes szám első személyű igealakokat a passzív jelen idejű igealakok váltották fel. Ebbe a

feladatba azok a példamondatok kerülhetnek be, amelyekben megtalálható legalább egy passzív alakban álló finit ige, azaz a morfoszintaktikai jegyek között mind a `Voice=Pass`, mind a `VerbForm=Fin` megtalálható. Ezt követően a mondatból eltávolítjuk a passzív alakot, és annak csak a lemmáját jelenítjük meg, hogy a tanuló ezt felhasználva be tudja helyettesíteni a megfelelő alakot. Ezen nyelvtani jelenséget összesen 2.092 finn mondaton lehet gyakorolni.

Magyar feladatok A magyar nyelvre szintén három feladattípus implementálása történt meg.

Az első feladatot az igekötőknek szenteltük. Az, hogy pontosan mely lexikai elemek számítanak az igekötők kategóriájába, vitatott téma (Kalivoda, 2021). Jelen alkalmazás 13 szót kezel igekötőként, de ezek listája bármikor bővíthető, illetve szűkíthető. A magyar igekötők mind igemódosító pozícióban (a finit ige előtt), mind az igetőt követően, posztverbálisan is megjelenhetnek. Az általunk használt `emtsv` eszköz dependenciaelemző modulja ugyan `PREVERB` címkével látja el az igekötőket, azonban erre csak akkor kerül sor, ha azok elválnak az igetőtől. A mi esetünkben a közvetlenül az ige előtt, preverbális helyzetben megjelenő igekötőkre is szükség van, így a mondatban előforduló karakterláncokra hagyatkozunk. Azon mondatokra lesz szükség ezen feladat megvalósításához, amelyek vagy valamely ige elején vagy önálló szóként tartalmazzák a következő szavak egyikét: `be`, `ki`, `le`, `fel`, `meg`, `el`, `át`, `bele`, `ide`, `oda`, `szét`, `össze`, `vissza`. Az igekötőt tartalmazó elemek közül 300 mondatot alaposan megvizsgálva azt állapíthattuk meg, hogy vannak bizonyos igék, amelyek ugyan valamely igekötővel kezdődnek, ezek mégis az igető részét képezik. Ilyen igékre példa a *beszél*, a *becsül*, a *felejt*, a *felel* vagy a *kiabál*. Ezeket az igéket összegyűjtöttük, és kizártuk az olyan mondatokat, amelyek ezek közül bármelyiket is tartalmazzák. Ezen szűréseket követően 5.227 példamondat maradt, ezek képezik részét az igekötőket feldolgozó feladatnak. A mondatokból eltávolítottuk az igekötőket, és a tanuló feladata az, hogy a 13 lehetséges szó közül kiválassza a mondatba leginkább illeszkedő elemet.

A határozott és határozatlan igeragozás közötti különbségtétel sok magyarul tanulónak okoz gondot. A tanuló olyan mondatok segítségével tudja ezen nyelvi jelenséget gyakorolni, amelyekben egy tranzitív ige található akár alanyi, akár tárgyas ragozásban. Ezt a morfoszintaktikai jegyek között megtalálható `Definite=Def` jegy–érték pár adja meg. Az lényegtelen, hogy egy tranzitív ige éppen milyen ragozásban fordul elő egy adott mondatban, hiszen a feladattal éppen az a célunk, hogy a tanulónak kelljen eldöntenie, mikor használunk alanyi és mikor tárgyas ragozást. Az alkalmazás ezen alegysége 5.830 mondatot tartalmaz. Itt szükség van a mondat alanyának számára és személyére, illetve az igeidőre ahhoz, hogy a helyes válasz megadható legyen.

A magyar birtokos szerkezet felépítése meglehetősen eltér sok más nyelv birtokos szerkezetétől. A magyar nem a birtokoson, hanem a birtokon tünteti fel a birtokos személyjelet. Ez a birtokot kifejező szó morfoszintaktikai jelei között úgy jelenik meg, hogy a `Number[psor]` és `Person[psor]` jegyek értékei adják meg a birtokos számát és személyét. Az adatbázisból azon mondatokat kell kigyűjteni,

amelyekben előfordul egy olyan token, amelynek morfoszintaktikai jegyei között ezen elemek megtalálhatók. Egyelőre a rendszerben csak a nominatívuszban álló birtokok (`Case=Nom`) szerepelnek, mivel a hangsúlyt a birtokos személyjel helyes feltüntetésére szeretnénk fektetni. Az adatok közül 4.896 példamondat illeszkedik erre a mintázatra. Ezen feladattípus során a mondatokból eltávolítottuk a birtokot kifejező szavakat, és a tanuló a birtok számát, valamint a birtokos számát és személyét is megkapja a behelyettesítendő szó lemmájával együtt, hiszen a magyar mondatokban előfordulhat üres névmási birtokos is (lásd 5. ábra). A birtokos számának és személyének feltüntetése nélkül a behelyettesítendő szó akármely számú és személyű birtokos esetén grammatikus mondatot eredményezne, ugyanígy a birtok egyes és többes számban is elfogadható lenne.

5. ábra: Üres névmási birtokos a magyarban.

A felhasználó válaszait minden feladattípus esetén egy külön táblában tároljuk, amely lehetőséget ad arra, hogy a jövőben az idegennyelv-elsajátítás során felmerülő problémákat a kutatók szabadon elemezhessék (a nyelvtanulók személyiségi jogait és a személyes adatok védelmét szem előtt tartva), illetve megvizsgálhassák az általunk összegyűjtött adatokat.

4.4. Kiértékelés

Minden fordítási pár, valamint mondat és mondatelemzés kézi ellenőrzésen esik át, mielőtt szabadon elérhetővé tennénk őket. Nagyon fontos ugyanis, hogy a nyelvtanulók ne tanuljanak meg hibás célnyelvi megfelelőket, nyelvtani szerkezeteket, elemzéseket. A kézi elemzés jelenleg is folyamatban van az összegyűjtött adatok száma miatt, azonban a mondatok egy kisebb halmazát kézzel ellenőrizve kimutatható a fent bemutatott finn és magyar nyelvtechnológiai eszközök pontossága. Az eddigi eredményeket a 4. táblázatban foglaltuk össze. Jól látható, hogy a finn mondatok esetén a lemmatizálás az esetek nagy részében hibás kimenethez vezet, amelyet az `omorfi` elemzőrendszer végez. Az 50 ellenőrzött finn mondat pontossága mindössze 30%. A magyar mondatokat az `emtsv` rendszer többnyire helyesen elemezte, ennek pontossága 68%.

A fennmaradó mondatok elemzése folyamatban van, alkalmazásunkat úgy alakítottuk ki, hogy a nyelvtanulók csak a kézi ellenőrzésen már átesett mondatokon gyakorolhatnak különböző nyelvtani jelenségeket.

5. Összegzés

A cikkben bemutattunk egy olyan nyelvtanuló alkalmazást, amely a finn és magyar szókincs építését, valamint a nyelvtan gyakorlását teszi lehetővé különböző

	finn (db)	% finn	magyar (db)	% magyar
Nem teljes mondat	2	4%	1	2%
Hibás lemmatizálás	28	56%	5	10%
Hibás morfológiai elemzés	3	6%	3	6%
Hibás dependenciaelemzés	2	4%	7	14%
Pontosság	15	30%	34	68%
Összes kiértékelt mondat	50	100%	50	100%

4. táblázat. A mondatok kiértékelése.

feladatokon keresztül. A feladattípusok könnyen bővíthetők, egyedül a nyelvi mintázatok pontos leírására van szükség ahhoz, hogy az adatbázisból új példamondatokat gyűjtsünk ki, és behelyettesítéses feladatot generáljunk belőlük automatikusan.

Ezen alkalmazás egyik hiányossága, hogy csak a rendszer által tárolt választ fogadja el egyetlen helyes megoldásként, az egyéb alternatívákat, amelyek szintén grammatikus mondatokhoz vezetnének, nem. A szókártyák esetében egy-egy definíció meghatározhat több fogalmat is, és van, hogy a meghatározás nem szolgál elegendő információval egy adott fogalom felismeréséhez. Amennyiben a felhasználó nem a rendszer által korábban ismertetett szót vagy kifejezést adja meg, az automatikus kiértékelés során válasza hibásnak lesz feltüntetve. Erre irányuló fejlesztéseket a projekt egy következő szakaszában fogunk végezni.

További probléma, hogy a feladatokban használt példamondatokat, illetve a szókártya alkalmazásban megjelenített új kifejezéseket nem osztályozzuk a tudásszinteknek megfelelően, ezáltal lehet, hogy a nyelvtanulók nem a saját nyelvi szintjüknek megfelelő feladatokat kapnak. A jövőben ezt úgy tervezzük megoldani, hogy a mondatokat és kifejezéseket automatikus módszerekkel besoroljuk a Közös Európai Referenciakeret (KER) egyes nyelvi szintjeibe, és csak a tanuló szintjének megfelelő elemeket jelenítünk meg.

Amint az itt bemutatott feladatok és az azokban használt mondatok kézi kiértékelése megtörténik, a webes alkalmazást szabadon elérhetővé kívánjuk tenni minden nyelvtanuló számára. A tanulók segítségével és a feladatokra adott válaszaikkal még pontosabb képet kaphatunk alkalmazásunk hasznosságáról, illetve azokról a pontokról, amelyek még javításra szorulnak a rendszerünkben. Kifejezetten érdekes lehet a jövőben az adatbázisban tárolt válaszok alapján kvantitatív vizsgálatokat folytatni, és esetlegesen olyan típushibákra bukkanni, amelyekről korábbi kutatások még nem tettek említést. Feltételezhető, hogy empirikusan alátámaszthatóvá válnak majd olyan elméleti kutatások, amelyek tárgya a finnugor nyelveket idegen nyelvként tanulók nyelvelsajátításával kapcsolatos nehézségeire irányulnak. Az adatok alapján kézzelfoghatóvá válik, hogy pontosan mely nyelvtani szerkezet jelenti a legnagyobb nehézséget a tanulók számára, és mely az, amelyet gond nélkül tudnak alkalmazni.

Hivatkozások

Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora. pp. 52–58. Association for Computational Linguistics, Sofia, Bulgaria (2013)

Antal, L.: A magyar esetrendszer. Nyelvtudományi Értekezések 29 (1961)

Antonsen, L., Huhmarniemi, S., Trosterud, T.: Interactive pedagogical programs based on constraint grammar. In: Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009). pp. 10–17 (2009)

Basoglu, E.B., Akdemir, O.: A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flash cards. Turkish Online Journal of Educational Technology-TOJET 9(3), 1–7 (2010)

Elgort, I.: Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. Language Testing 30(2), 253–272 (2013)

Hakulinen, A., Vilkuna, M., Korhonen, R., Kovisto, V., Heinonen, T.R., Alho, I.: Iso suomen kielioppi [Nagy finn nyelvtan]. In: SKS:n toimituksia 950. Suomalaisen Kirjallisuuden Seura, Helsinki (2004)

Hämäläinen, M.: UralicNLP: An NLP library for Uralic languages. Journal of Open Source Software 4(37), 1345 (2019)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M.: One format to rule them all – The emtsv pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 155–165. Association for Computational Linguistics, Florence, Italy (2019)

Jo, G.: English Vocabulary Learning with Wordlists vs. Flashcards; L1 Definitions vs. L2 Definitions; Abstract Words vs. Concrete Words. Culminating Projects in English 132 (2018)

Kalivoda, Á.: Igekötős szerkezetek a magyarban. Ph.D.-értekezés, Pázmány Péter Katolikus Egyetem (2021)

Karlsson, F., Chesterman, A.: Finnish: An Essential Grammar. Routledge (2008)

Katinskaia, A., Nouri, J., Yangarber, R.: Revita: a system for language learning and supporting endangered languages. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition. pp. 27–35 (2017)

Kiefer, F.: The Cases of Hungarian Nouns. Acta Linguistica Academiae Scientiarum Hungaricae 37(1/4), 93–101 (1987)

Kiefer, F.: Magyar nyelv. Akadémiai Kiadó (2006)

Kiefer, F.: Strukturális magyar nyelvtan 3. kötet: Morfológia. Akadémiai Kiadó (2018)

Kilickaya, F., Krajka, J.: Comparative usefulness of online and traditional vocabulary learning. Turkish Online Journal of Educational Technology-TOJET 9(2), 55–63 (2010)

Laufer, B., Elder, C., Hill, K., Congdon, P.: Size and strength: Do we need both to measure vocabulary knowledge? Language testing 21(2), 202–226 (2004)

Lindén, K., Carlson, L.: FinnWordNet–Finnish WordNet by Translation. LexicoNordica–Nordic Journal of Lexicography 17, 119–140 (2010)

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)

Máté, J.: A magyar nyelv elsajátításának nehézségei a finn anyanyelvű tanulók szempontjából. Hungarologische Beiträge 12, 91–112 (1999)

Pirinen, T.A.: Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics 28, 381–393 (2015)

Simon, V.: Ensiapu – Elsősegély: Módszertani segédanyag a finn nyelv oktatásához. Eötvös Loránd Tudományegyetem, Budapest (2015)

Tiedemann, J., Nygaard, L.: The OPUS Corpus - Parallel and Free. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal (2004)

Tompa, J.: A Mai magyar nyelv rendszere: leíró nyelvtan. Akadémiai Kiadó, Budapest (1961)

Uibo, H., Pruulmann-Vengerfeldt, J., Rueter, J., Iva, S.: Oahpa! Õpi! Opiq! Developing free online programs for learning Estonian and Võro. In: Proceedings of the fourth workshop on NLP for computer-assisted language learning. pp. 51–64 (2015)