Finn–magyar fordítási párok kinyerése automatikus módszerekkel

Ferenczi Zsanett

PPKE BTK Nyelvtudományi Doktori Iskola

ferenczizsani@gmail.com

Kivonat: A kétnyelvű szótárak a nyelvtanulás fontos eszközei, kontrasztív kutatások számára nagyszerű alapanyagot nyújtanak, ezen kívül azonban a számítógép által is értelmezhető kétnyelvű lexikonokra különböző nyelvtechnológiai eszközök (például statisztikai gépi fordítórendszerek) fejlesztésekor is szükségünk lehet. Ilyen szólisták manuálisan történő összeállítása ugyanakkor munkaerő- és időigényes folyamat. Ez a munkafolyamat automatizálással felgyorsítható, de ezen módszerek egyik gyengéje, hogy nem lehet velük tökéletes pontosságot elérni. Jelen kutatásban a szótárkészítés egyik első lépését mutatom be: egy olyan finn–magyar szavakat és többszavas kifejezéseket tartalmazó kétnyelvű lista összeállítását, melyet különböző lexikai erőforrásokból automatikus módszerekkel sikerült kinyernem, s mely két nyelv címszójegyzékeként is fog szolgálni egy elkészítendő online szótárhoz. Ismertetek három új algoritmust, melyek a már létező eljárások eredményein túl további szópárok kinyeréséhez vezettek. Ezen metódusok segítségével több százezer fordítási párt sikerült összegyűjtenem.

https://doi.org/10.18135/Alknyelvdok.2021.15.6
In: Gráczi Tekla Etelka – Ludányi Zsófia (szerk.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2021.* Budapest: Nyelvtudományi Kutatóközpont. 2021. (Sorozatszerkesztő: Váradi Tamás.)
ISBN 978-963-9074-94-1.

1. Bevezetés

A kétnyelvű szótárak az idegen nyelvet tanulók, fordítók számára nagy segítséget nyújtanak, míg kétnyelvű szólisták bizonyos nyelvtechnológiai eszközök létrehozásakor, illetve ezek minőségének javításakor is elengedhetetlenek.

A nyomtatott szótárak után az 1950-es évek végén, az 1960-as évek elején megkezdődtek az elektronikusszótár-építő munkálatok, majd az 1990-es évek végétől kezdve megjelentek az online szótárak. Az internet, valamint a hordozható eszközök (laptop, okostelefon, táblagép) szélesebb körben való elterjedésével ezen lexikai adatbázisok egyre nagyobb figyelmet kaptak. Gaál (2016) felmérése alapján is jól látható, hogy a szótárhasználók főleg az online szótárakat részesítik előnyben, a kizárólag papírszótárt használók száma egyre inkább csökkenni látszik. Az elmúlt 60 évben az elektronikusszótárkészítés is nagy változásokon esett át: a legelső elektronikus szótárak mindössze a szólisták ábécérendbe sorolásához, az adatok rendezéséhez és a bevitt adatok ellenőrzéséhez vették igénybe a számítógép (és akkoriban még a számítógép kezeléséhez értő szakemberek) segítségét, meggyorsítva ezzel a szótárírás folyamatát. Manapság egy-egy szótáríró rendszer már szerves részét képezi a lexikográfusok eszköztárának. A szótár kiinduló anyagát cédulák helyett a korpuszok biztosítják. Az elektronikus szótár előnyei közé tartozik a fent említetteken kívül az is, hogy a szótár és a szócikkek mérete, terjedelme nem limitált, folyamatosan lehet frissíteni és újabb szócikkekkel bővíteni, valamint a felhasználók számára is sok kedvező funkciót rejt némely modern szótár (intelligens keresés, testreszabhatóság, hivatkozások használata a szócikkek közötti gyors navigáláshoz stb.). Ezek készülhetnek a hagyományos módon, szakértők által kidolgozva minden egyes szócikket, esetleg automatikus módszerekkel kinyerve a szócikkek tartalmát, vagy alapulhatnak a közreműködni vágyó felhasználók tudásán, amennyiben Wiki-alapú, közösség által szerkesztett szótárról van szó.

A legelső finn–magyar szótár Szinnyei József nevéhez fűződik, aki 1884-ben adta ki körülbelül 15000 szócikket tartalmazó szótárát (Szinnyei, 1884). Maticsák és Laihonen (2011) kritikája sorra veszi a finn–magyar és magyar–finn kétnyelvű nyomtatott szótárakat. Megállapítják, hogy viszonylag sok ilyen erőforrás létezik, és ezek minősége is meglepően jó. Ennek ellenére azt is hangsúlyozzák összefoglalójukban, hogy a szókincs állandó megújulása miatt szükség van a szótárak folyamatos frissítésére, újabb kiadások megjelentetésére. A legújabb finn–magyar szótár 2015-ben jelent meg, ez az első olyan finn és magyar nyelveket feldolgozó szótár, melyet Finnországban adtak ki, főszerkesztője Ulla-Maija Forsberg (Forsberg és mtsai., 2015). Ebben a szótárban a szócikkek száma meghaladja a 41500-at, mely így majdnem háromszor annyi, mint amennyi a legelső, Szinnyei (1884) által írt XIX. századi szótárban található. Ez a referenciamű azonban több mint tíz éven át készült, így felmerülhet a kérdés: mennyire lehet ezt a munkafolyamatot nyers fordítási párok kinyerésével felgyorsítani automatikus módszereket bevetve?

Az interneten is sok finn–magyar és magyar–finn szótárt találhatunk, ezek többsége ingyenesen használható. Az egyik legnagyobb és legtöbb nyelvet (köztük a finnt és magyart is) feldolgozó szótár a Wiktionary, mely egy közösség által szerkesztett szótár. Ennek nyelvi anyagát dolgozza fel sok más finn–magyar–finn online szótárprojekt (`finnhun.com`; `sanakirja.org`; `ilmainensanakirja.fi`), így ezek csak megjelenésüket és elérési útvonalukat tekintve különböznek a Wiktionary oldalától. Létezik egy olyan internetes szótár is erre és még sok másik nyelvpárra, amelyet nyelvészek manuálisan állítottak össze (`dict.com`), itt azonban – az összeállítás módja miatt, érthető okokból – csak 16300 (finn–magyar irányban), illetve 17400 (magyar–finn irányban) szócikkhez férhetünk hozzá ingyenesen. Ezen online szótárak egyik nagy hiányossága, hogy a célnyelvi jelentéseket nem pontosítják, és ez homonímia, illetve poliszémia esetén a szótár használhatatlanságához vezet. A szó jelentésének, használatának jobb megértését

segítő példamondatok is csak néhány szótárban tűnnek fel, és ezeket többnyire automatikusan emelik ki párhuzamos szövegkorpuszokból. Ezeket a példamondatpárokat a weboldal alján ömlesztve sorolják fel, mely áttekinthetetlen, zsúfolt megjelenéshez vezet, megszegve a jó minőségű online szótárak egyik, Gaál (2012) által meghatározott alapkövetelményét. A közösség által szerkesztett online szótárakkal kapcsolatban Gaál (2016) azt is megfigyelte, hogy a felhasználók megbízhatatlanabbnak tartják, mint szakértők, intézmények, szótárkiadó cégek által szerkesztett párjaikat. Ennek ellenére az ingyenes tartalom a leginkább preferált opció az online szótárakat használók körében, még akkor is, ha ez az oldalon hirdetések, reklámok jelenlétét vonja maga után.

Az interneten ingyenesen elérhető két- és többnyelvű szótárakat elemezve tehát kijelenthető, hogy egy a felhasználók igényeit szem előtt tartó, jó minőségű finn–magyar–finn internetes szótár még várat magára. Erre szeretnék megoldást kínálni egy olyan ingyenes online szótár elkészítésével, mely két fontos előnyt próbál ötvözni: a gyorsaságot és a pontosságot. Ezt hibrid módszerrel valósítom meg: a szótárba foglalandó címszavakat, a szócikkek alapjait és egyéb információkat (például példamondatokat, definíciókat) automatikus módszerekkel nyerem ki (melynek lépéseit jelen dolgozatban mutatom be), majd ezen nyers adatok kézi ellenőrzésen és kiegészítésen esnek át, így biztosítva a nyelvi adatok helyességét és a szótár teljességét.

2. Módszertan

Az elmúlt pár évtizedben számos megoldás született a kétnyelvű szótárépítés automatikus kivitelezésére. Eleinte főleg párhuzamos korpuszokból próbáltak kétnyelvű adatokat kinyerni (Kumano és Hirakawa, 1994; Wu és Xia, 1994), de mivel nem minden nyelvpár esetén lehetett párhuzamos korpuszra szert tenni, a kutatók egyre

több olyan módszerrel kezdtek kísérletezni, melyek az összevethető korpuszokat vették alapul (Fung, 1995; Rapp, 1995; Fung és Yee, 1998). Ezeken túlmenően léteznek olyan megoldások is, melyek alternatív módszerekkel kapcsolnak össze többnyelvű erőforrásokat, és így eredményeznek fordítási párokat.

Kutatásomban főleg ez utóbbi módon próbáltam többféle erőforrásból is kétnyelvű szó- és kifejezéspárokat kinyerni a finn–magyar nyelvpárra, valamint az OPUS-ban (Tiedemann és Nygaard, 2004) megtalálható filmfeliratok és más párhuzamosnak tekinthető korpuszok (Európai Parlament dokumentumai, szoftverdokumentációk) felhasználásával készült szópárlisták elemeivel is kiegészítettem a nyers fordítási párok halmazát.

A következőkben bemutatom a felhasznált erőforrásokat, majd ismertetem az alkalmazott, illetve létrehozott algoritmusok lépéseit, eredményeit és elérhetőségét.

2.1. Wiktionary

A Wiktionary egy többnyelvű, webalapú szótárprojekt, melynek célja, hogy minél több nyelv minél több szavát tartalmazza. 2020 januárjában a Wiktionary 174 nyelven volt elérhető. A Wiktionary-kiadás nyelve határozza meg a szótár célnyelvét, a forrásnyelv azonban nem meghatározott: bármely nyelvű verzióban kereshetünk bármilyen nyelvű szócikkre. E között a 174 nyelv között megtalálható a finn (Wikisanakirja) és a magyar (Wikiszótár) is. A Wiktionary-szócikkek felépítését és minimális követelményeit az egyes kiadások eltérően határozzák meg, így a különböző nyelvű verziókhoz különböző elemzőkre van szükségünk. A szócikkek egyik fontos alkotóeleme a forrásnyelv megjelölése, mivel egy-egy címszó több különböző nyelvhez is tartozhat (például tag). Általában megjelenik az adott szó szófaja vagy a többszavas kifejezés típusa (közmondás, kifejezés stb.), illetve a célnyelven (az adott Wiktionary-kiadás nyelvén) való jelentés, fordítás is. A forrásnyelv olykor megegyezik a célnyelvvel,

ilyenkor a Wiktionary értelmező szótárként működik, és a célnyelvi oldalon fordítás helyett definíciókkal, szinonimákkal, körülírással találkozhatunk. Ezen esetek többségében egy fordítási tábla is tartozik a szócikkhez, amelyben az adott kifejezés fordításai vannak összegyűjtve minden olyan nyelven, amelyhez tartozik szócikk az adott Wiktionary-kiadásban.

2.2. WordNet

A WordNet egy nagy lexikai adatbázis, mely főneveket, igéket, mellékneveket és határozószókat sorol be synsetekbe (szinonimahalmazokba), jelentésük alapján. Ezeket a synseteket alapvető lexikális szemantikai relációk kötik össze, melyek között fellelhető a hiperonímia, az antonímia és a metonímia is. Az első WordNet az angol nyelvre készült (Miller és mtsai., 1990), majd több nyelvre lefordították ezt az adatbázist, köztük magyarra (Miháltz és mtsai., 2008) és finnre (Lindén és Carlson, 2010) is. A magyar WordNet 42288 synsetet és 50238 különböző szót, kifejezést foglal magába, míg a finn verzió a fordítást követően kibővült egyéb szinonimahalmazokkal is, és így 120449 synsetet és 140515 szót tartalmaz. A számok alapján jól látható, hogy az egyes synsetekben található szavak, kifejezések száma átlagosan több mint 1 (1,188 a magyar, illetve 1,167 a finn esetében).

A finn WordNet az angol Princeton WordNet (PWN) 3.0 verzióján alapul (első körben ennek fordításaként jött létre). A magyar WordNet ugyan a BalkaNet fogalmi halmazait vette kiindulópontul, mely a PWN 2.0-s verziójára épít, ezeket azonban később leképezték a PWN 3.0 verziójában szereplő synset offsetekre (a szinonimahalmazokat egyedi módon azonosító, nyolc számjegyből álló kulcsokra), így ezen azonosítók alapján össze lehet kapcsolni a két adatbázist, és meg lehet egymásnak feleltetni a fogalmi halmazokat. Az 1. ábrán látható egy magyar synset (XML-formátumban) és a finn WordNetben ugyanennek az offsetnek megfelelő három találat

(TSV-formátumban).

A két WordNet felépítésének különbségét ez az ábra jól szemlélteti. Az egyedi azonosítók szerkezetének eltérése is szembetűnő: míg a magyar offset `ENG30-00313502`-n alakú, addig a finn offsetek felépítése a következő: `fi:n00313502`. A WordNetek különböző struktúrája miatt olyan programra van szükségünk, mely ezen eltéréseket képes kezelni és a szükséges összekötéseket elvégezni.

```
<SYNSET>
  <ID>ENG20-00299650-n</ID>
  <ID3>ENG30-00313502-n</ID3>
  <POS>n</POS>
 <SYNONYM>
    <LITERAL>űrrepülés<SENSE>1</SENSE></LITERAL>
    <LITERAL>űrutazás<SENSE>1</SENSE></LITERAL></SYNONYM>
  <ILR>ENG20-00298868-n<TYPE>hypernym</TYPE></ILR>
  <DEF>Űrhajóval az űrben végzett repülés.</DEF>
  <USAGE>Az idei második űrrepülés sikeresen zajlott.</USAGE>
  <STAMP>Rendszergazda 2008/02/07</STAMP>
  <DOMAIN>factotum</DOMAIN>
  <SUMO>Transportation<TYPE>+</TYPE></SUMO>
  <EKSZ>űrrepülés_1_1<TYPE>=</TYPE></EKSZ>
</SYNSET>
```

```
fi:n00313502  avaruusmatkailu  en-3.0:n00313502  spacefaring  synonym
fi:n00313502  avaruuslento     en-3.0:n00313502  spaceflight  synonym
fi:n00313502  avaruusmatka     en-3.0:n00313502  space travel synonym
```

1. ábra. Példa a magyar és finn WordNet synsetekre.

3. OPUS

Az OPUS egy szabadon hozzáférhető, mondatok szintjén párhuzamosított korpuszokat tartalmazó gyűjtemény (Tiedemann és Ny-

gaard, 2004), mely többek között filmfeliratokat, szoftverek dokumentációját és az Európai Parlament dokumentumait dolgozza fel. Aulamo és munkatársai elemzése alapján (Aulamo és mtsai., 2020) összesen 57 korpuszt tartalmaz, és több mint 700 nyelvet és nyelvváltozatot fed le. Egyes korpuszok és nyelvpárok esetében szószintű párhuzamosításokat (Tiedemann és Nygaard, 2004) is letölthetünk az oldalról. A finn–magyar pár 23 különböző forrásból származó párhuzamos korpusszal rendelkezik, ezek közül 12 esetben találhatunk kétnyelvű szópárokat tartalmazó listákat is. Mivel a szópárok a nyers korpuszokból lettek kinyerve, a szavak többnyire nem szótári alakjukban szerepelnek. A párhuzamosítást megelőző tokenizáció miatt többszavas kifejezéseket nem találhatunk ebben az erőforrásban. Az 1. táblázatban látható egy kisebb minta az OPUS-ról letöltött nyers szólistákból, mely azt szemlélteti, hogy lemmatizálás nélkül ezen 8 szópárból egyik sem vezetne a helyes fordítási párhoz, mely a poika szóhoz a fiú szót rendeli. A szavak lemmatizálását tehát mindenképpen érdemes lenne elvégezni, mely pontosabb fordítási párokat eredményezne.

Poika	A
Poikaa	a
poikaa	fia
Poikaa	Fia
poikaa	fiat
poikaa	fiát
poikaa	fiú
poikaa	fiút

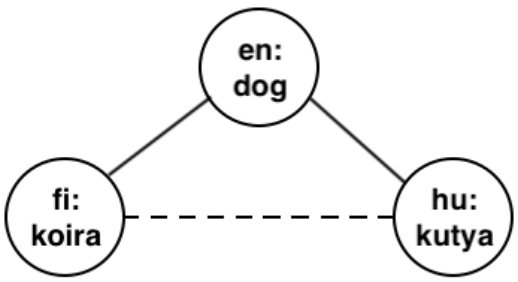
1. táblázat. Párhuzamosított szópárok az OPUS-ról

Mivel a fent ismertetett erőforrások alapvető felépítésükben különböznek, eltérő módszereket alkalmaztam a szópárok kinyeréséhez.

3.1. wikt2dict

A `wikt2dict` eszközt Ács és mtsai. (2013) alkották meg, mely a Wiktionary fordítási tábláiból nyer ki fordítási párokat. Ezt az algoritmust két módon lehet futtatni: az extract és a triangulate parancsokat használva. Az extract két (vagy több) nyelvet kap meg paraméterként, és az egyik nyelv Wiktionaryjéből nyeri ki az összes olyan szót és kifejezést, amely a másik nyelven megjelenik a fordítási táblákban. Ez olyan szócikkeket elemez, melyek az adott Wiktionary-kiadás nyelvét is felsorolják a forrásnyelvek között. Ezt a szkriptet a finn és magyar nyelvű Wiktionary-verziókon futtattam le.

A triangulate algoritmusa azon a feltételezésen alapul, hogy ha egy fordítási táblában található két tetszőleges nyelvű kifejezés ugyanazon szónak, szókapcsolatnak a fordítása, akkor azok egymás fordításainak is tekinthetők. Ezen módszer alapötletét szemlélteti a 2. ábra, ahol a finn és magyar szavak között az angol dog szócikk fordítási táblájának segítségével állapítható meg a kapcsolat. Ehhez a szkripthez tehát három nyelvkódra van szükség, az eredmény pedig egy fordítási hármasokat tartalmazó lista, melyből a harmadik nyelvet figyelmen kívül hagyva könnyen finn–magyar szópárokhoz juthatunk. A legtöbb szócikket (és legtöbb információt) tartalmazó Wiktionary az angol, ezért a finn és magyar mellé ezt vettem fel harmadik nyelvként a háromszögeléshez.



2. ábra. Új fordítási párok kinyerése egy harmadik nyelv segítségével

3.2. Wiktionary Parser

Ahogy korábban említettem, a Wiktionary nyelve az adott szótár célnyelveként funkcionál, így nem csak a fordítási táblákat elemezve lehet szópárokat kinyerni egy adott nyelvpárra, hanem a szócikk fő tartalmi blokkjából is. Itt többek között a forrásnyelvek, a szó szófajai és a célnyelvi megfelelők jelennek meg. Ahhoz, hogy fordítási párokat nyerhessek ki a finn és a magyar Wiktionaryből, egy olyan szkriptre volt szükségem, mely ezeket a verziókat képes elemezni, bejárni és így kigyűjteni a releváns tartalmat. Ennek feldolgozására létrehoztam egy algoritmust, mely a Wikisanakirjában a magyar, a Wikiszótárban pedig a finn szócikkeket keresi meg, nyeri ki, és tárolja el a szükséges információkat. A szócikkek forrásnyelvi oldalán állhat többszavas kifejezés, valamint a célnyelvi oldalon is megjelenhet körülírás, felsorolás vagy idiomatikus kifejezés, ezeket mind megőrzi a létrehozott szkript. Minden szócikk esetén eltárolom a fordítási párt, illetve a hozzájuk tartozó szófaji információt.

Emellett a Wiktionary értelmező szótár funkcióját is kihasználom, a Wikiszótárban a magyar nyelvű szócikkekhez, a Wikisanakirjában pedig a finn nyelvű szócikkekhez gyűjtöm ki a definíciókat és a példamondatokat, a címszóra vonatkozó szófaji információval együtt, hogy a homonim címszavak esetén ez segítse a jelentés beazonosítását. A kinyert információkat szemléltető példák a 2. táblázatban találhatók.

A létrehozott algoritmus forráskódja a Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA) licenc alatt szabadon elérhető és felhasználható (https://github.com/ferenczizsani/wiktionary_parser).

3.3. WordNet Connector

A finn és magyar nyelvű WordNetek összekapcsolása a Princeton WordNet 3.0 verziójának synset offsetjei által lehetséges. Az egyik nyelv synsetjeiben szereplő kifejezéseket a másik nyelv ugyanazon

synset offsethez tartozó összes szavával összepárosítva fordítási párok halmazaihoz jutunk, melyek esetében a szófaj is ismert. Ezeket a synset offsetekben jelöli mindkét nyelvű WordNet, egy-egy karaktert rendelve a négy szófaji kategóriához (n – főnevek, v – igék, a – melléknevek, r (finn) és b (magyar) – határozószók). A szkript először végigjárja a finn, majd a magyar WordNetet, és kigyűjti az egyes szófajokhoz és offsetekhez tartozó szavakat, kifejezéseket úgy, hogy azok ne ismétlődjenek. A korábban bemutatott WordNet-részletek (1. ábra) összekapcsolása után a 3. táblázatban látható fordítási párokat kapjuk.

fordítások			
ikkuna		*ablak*	NOUN
hyvää yötä		*jó éjszakát*	INTJ
példamondatok			
finn	NOUN	*kahvi*	*Tuo kaupasta kahvia!*
magyar	NOUN	*réz*	*Ebben a kis faluban rezet bányásznak.*
definíciók			
finn	ADV	*toki*	*tietenkin, tietysti, miksipä ei*
magyar	NOUN	*vár*	*fallal körülvett erőd*
magyar	VERB	vár	*valahol marad és nem csinál semmit mindaddig, amíg valaki meg nem jön vagy valami nem történik*

2. táblázat. Példa a Wiktionary Parser által kinyert adatokra

A magyar WordNetben találhatunk definíciókat és példamondatokat is az egyes szinonimahalmazokhoz. A definíciót a <DEF>, míg a használatot szemléltető példamondatot a <USAGE> tagekkel jelölik (ahogyan azt már az 1. ábrán is láthattuk). Ezeket a szkript automatikusan kinyeri, és a szinonimahalmazban található valamennyi szóhoz, kifejezéshez hozzárendeli azokat, a szófaji kategóriát és a synset offsetet is megőrizve (lásd 4. táblázat). A kézi

avaruusmatkailu	űrutazás	NOUN
avaruusmatkailu	űrrepülés	NOUN
avaruuslento	űrutazás	NOUN
avaruuslento	űrrepülés	NOUN
avaruusmatka	űrutazás	NOUN
avaruusmatka	űrrepülés	NOUN

3. táblázat. Példa a WordNet Connector által kinyert fordítási párokra

ellenőrzés során el kell majd távolítani azon sorokat, amelyek esetén a példamondat valójában nem tartalmazza a megadott lemmát, mivel egy szinonimahalmazhoz (amely egynél több szót, kifejezést is tartalmazhat) csak egy ilyen mondat tartozik. A definíciók esetén ilyen szűrésre nincs szükség, mivel azok a szinonimahalmaz minden elemére érvényesek.

definíciók			
NOUN	00313502	űrutazás	Űrhajóval az űrben végzett repülés.
NOUN	00313502	űrrepülés	Űrhajóval az űrben végzett repülés.
példamondatok			
NOUN	00313502	űrutazás	Az idei második űrrepülés sikeresen zajlott.
NOUN	00313502	űrrepülés	Az idei második űrrepülés sikeresen zajlott.

4. táblázat. Példamondatok és definíciók a magyar WordNetből

A WordNeteket összekapcsoló szkript a GitHubon elérhető és szabadon felhasználható Creative Commons Attribution-ShareAlike 4.0 licenc forrásmegjelölés mellett (https://github.com/ferenczizsani/connect_wordnets).

3.4. OPUS Extractor

Az OPUS-ban található 12 párhuzamos szólistát letöltöttem a finn–magyar nyelvpárra, majd a kapott fájlokból a megfelelő mezők kiszűrése után egységes formátumra hoztam az adatokat. A felhasznált korpuszok jellege és a lemmatizálás hiánya miatt látszólag ez a metódus eredményezi a legpontatlanabb szópárokat. Két fő hibatípust lehet megfigyelni: a szoftverdokumentációk és a filmfeliratok olyan szavakat is tartalmaznak, melyek valójában sem a finn, sem a magyar nyelvben nem találhatók meg (pl. csak cirill karaktereket tartalmazó szavak), valamint a lemmatizálás hiánya miatt egy adott szóalakhoz a másik nyelvben sok különböző esetű (de ugyanazon lemmához tartozó) szóalak van rendelve. Ennek elkerülése érdekében a kinyerést követően lemmatizálást végeztem az egyes nyelvek szólistáin. A finn esetében ezt az `omorfi` (Pirinen, 2015), míg a magyar nyelvre az `emMorph` és az `emMorph2UD` (Indig és mtsai., 2019) eszközök segítségével oldottam meg. Ennek során sok más hibásan összepárosított fordítást is ki lehetett zárni azon feltételezés segítségével, mely szerint a fordítási párban szereplő szavak szófaji kategóriáinak azonosnak kell lennie. Így az 1. táblázatban látható szópárok száma mindössze kettőre csökkent *(poika – fia* és *poika – fiú)*.

Az algoritmusra, amely letölti és kigyűjti a releváns tartalmat az OPUS szólistáiból, a Creative Commons Attribution-ShareAlike 4.0 licenc feltételei érvényesek (`https://github.com/ferenczi-zsani/opus_extractor`).

3.5. Szófajok

A szócikkek egyik fontos alkotóeleme a szófaji címke. Ez sokszor segítséget nyújthat a szótárhasználó számára, főleg ha egy forrásnyelvi szó több szófaji kategóriához is tartozik, és ez nehezíti a helyes célnyelvi szó kiválasztását. Szófajok kinyerése történhet morfológiai elemzőkkel, amennyiben a felhasznált erőforrás nem tartalmazza ezt az információt. A fenti megoldások közül több is kinyert szófajt a

szavakhoz: a Wiktionary esetén része a szócikkeknek, a WordNet pedig a synset offsetben jeleníti meg a megfelelő szófaji kategória rövidítését. Ebben az esetben elegendő ezt az információt kinyerni és hozzárendelni a megadott szópárhoz. A fentebb felvázolt algoritmusok közül a `wikt2dict` és az `OPUS Extractor` azonban nem rendel szófajt a fordítási párok elemeihez. Ebben az esetben ugyanazon morfológiai elemzőket kellett bevetni, melyeket az OPUS lemmatizálásához is használtam. A Wiktionary és a WordNet azonban eltérő szófaji címkékkel dolgoznak, így az is egy fontos feladat volt, hogy ezeket egységes kategóriákra képezzem le. A Universal Dependencies szófaji címkekészletét (Nivre és mtsai., 2016) használtam minden esetben, ez összesen 17 címkét foglal magába.

4. Eredmények

4.1. Definíciók, példamondatok

A Wiktionary és a WordNet bizonyos esetekben tartalmaz a lemmákhoz kapcsolódó definíciókat és példamondatokat is. Az általam írt szkriptek minden ilyen esetben kinyerik a megfelelő tartalmakat, mivel ezekre egy online szótár építésekor szükség lehet. Segítséget nyújt a szótárhasználóknak a homonim címszavak különböző jelentéseinek megkülönböztetésében és egy-egy ismeretlen szó könnyebb megértésében. Az összegyűjtött példamondatok és definíciók címszavakkal alkotott kapcsolatainak ellenőrzése, esetleges törlése a kézi validáláskor fog megtörténni.

A magyar WordNeten található definíciók száma 22944, a magyar Wiktionaryről 26389, míg a finn Wiktionaryről 93632 különböző definíciót sikerült kinyerni. A két forrásból származó magyar nyelvű definíciókat össze lehet fésülni, az egymástól legalább egy karakterben eltérő definíciók száma összesen így 49293, azaz mindössze 40 definíció közös a két erőforrásban.

A példamondatok száma a magyar WordNeten 19024, a ma-

gyar Wiktionaryn mindössze 814, míg a finn Wiktionary szócikkei összesen 28837 példamondatot tartalmaztak. A Wiktionaryből és a WordNetről származó magyar adatokat automatikusan ellenőrizve és összefésülve kijelenthető, hogy a két példamondathalmaz kiegészíti egymást, ugyanis egyetlen olyan elem sincs, amely azonos lenne a két példamondathalmazban. Ezen eredményeket az 5. táblázat foglalja össze.

kinyert adat	**felhasznált erőforrás**	**darabszám**
magyar definíciók	WordNet és Wiktionary	49293
finn definíciók	Wiktionary	93632
magyar példamondatok	WordNet és Wiktionary	19838
finn példamondatok	Wiktionary	28837

5. táblázat. A Wiktionaryből és WordNetről kinyert definíciók és példamondatok száma

4.2. Szópárok

A fent bemutatott öt módszer eredményének összevetését a 6. táblázatban számokkal ábrázolom. Ezek alapján megállapítható, hogy az OPUS hozta a legtöbb szópárt, még a lemmatizálást követően is, ezek száma 378303. Ennek egyik oka az, hogy a párhuzamosított szólisták automatikus módszerekkel lettek kinyerve, ellentétben a WordNettel, melyet szakértők állítottak össze manuálisan, valamint a Wiktionaryvel, mely szintén emberi munka eredménye, és megvannak a minimális követelményei a szócikkek tekintetében. A `wikt2dict` háromszögelési módszere összesen 8535 találathoz vezetett, az extract parancs pedig közel tizenháromezer fordítást talált meg. A WordNetek közötti összekapcsolás 98883 pár létrehozását tette lehetővé, mely önmagában duplája a legfrissebb finn–magyar nyomtatott szótárnak.

módszer	kinyert fordítások száma
wikt2dict extract	12699
wikt2dict triangulate	8535
Wiktionary Parser	9370
WordNet Connector	98883
OPUS (lemmatizálás után)	378303
összesen	**473265**

6. táblázat. A Wiktionaryből és WordNetről kinyert definíciók és példamondatok száma

A különböző módszerek által kinyert fordításokat szintén automatikus módszerekkel fűztem össze, és eltávolítottam a duplikátumokat. Így összesen 473265 különböző finn–magyar szó- és kifejezéspárt sikerült összegyűjtenem. Ez több mint tízszerese a 2015-ben megjelent Suomi–unkari-sanakirja szócikkszámának, illetve huszonhétszerese a nyelvészek által összeállított online finn–magyar–finn szótárnak (dict.com).

5. Következtetések

Az itt bemutatott módokon sikerült több mint négyszázezer egyedi fordítási párt kinyernem a finn–magyar nyelvpárra, kizárólag automatikus módszereket használva. Mivel az egyes algoritmusok futási ideje nem tart tovább mindössze néhány percnél, elmondható, hogy az automatikus szótárépítő megoldások jóval gyorsabban több szó- és kifejezéspár kinyeréséhez vezetnek, mint az manuális módszerekkel elérhető lenne.

Természetesen a nyers fordítási párok önállóan nem használhatók nyelvtanulóknak szánt kétnyelvű szótárként, az egyes szócikkek tartalmának bővítéséhez egyéb kiegészítő információkra is szükség van, mint például szófaji kategória, példamondat, definíció, stílusminősítés és ragozási kategória. Ezen adatok többségét szintén automati-

kus módszerekkel nyertem ki, azonban jobb pontosság elérése és a hiányzó információk pótlása érdekében az így összeállított szócikkkezdemények kézi validáláson fognak átesni. A kézi ellenőrzés után a dolgozatban leírt módszerek kiértékelése is megtörténhet, és megvizsgálhatom, mely metódussal kapom a legpontosabb fordítási párokat.

Mivel az automatikus szótárépítő metódusok alapvetően nyelvfüggetlenek, a jövőben tervezem több finnugor nyelvre is kiterjeszteni és felhasználni ugyanezen módszereket.

Források

https://www.wiktionary.org. (A letöltés ideje: 2021. 03. 10.)

Finn Wikisanakirja: https://www.fi.wiktionary.org. (A letöltés ideje: 2021. 03. 10.)

Magyar Wikiszótár: https://www.hu.wiktionary.org. (A letöltés ideje: 2021. 03. 10.)

Finnhun.com: https://finnhun.com (A letöltés ideje: 2021. 03. 10.)

Sanakirja.org: https://sanakirja.org. (A letöltés ideje: 2021. 03. 10.)

Ilmainen Sanakirja: https://ilmainensanakirja.fi/suomi-unkari. (A letöltés ideje: 2021. 03. 10.)

Dict.com: https://dict.com (A letöltés ideje: 2021. 03. 10.)

Wiktionary Parser: https://github.com/ferenczizsani/wiktionary_parser. (A letöltés ideje: 2021. 03. 10.)

WordNet Connector: https://github.com/ferenczizsani/connect_wordnets. (A letöltés ideje: 2021. 03. 10.)

OPUS Extractor: https://github.com/ferenczizsani/opus_extractor. (A letöltés ideje: 2021. 03. 10.)

Universal Dependencies szófaji címkekészlete: https://univer-

saldependencies.org/u/pos/. (A letöltés ideje: 2021. 03. 10.)

Irodalom

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

Aulamo, M., Sulubacak, U., Virpioja, S. és Tiedemann, J. (2020). Opustools and parallel corpus diagnostics. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. és Piperidis, S. (szerk.): *Proceedings of the 12th Language Resources and Evaluation Conference*, 3782–3789, Marseille. European Language Resources Association.

Forsberg, U., Kovács., M., Kovács, O., Manner, S., Markus, K. és Vecsernyés, I. (2015). *Finn–magyar szótár*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In Yarowsky, D. és Church, K. (szerk.): *Proceedings of the Third Workshop on Very Large Corpora*, 173–183, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Fung, P. és Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In Yarowsky, D. és Church, K. (szerk.): *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1.*, 414–420, Montréal. Association for Computational Lingustics.

Gaál, P. (2012). Szempontrendszer online szótárak minősítéséhez. *Magyar Terminológia*, 5(2):225–250.

Gaál, P. (2016). Onlineszótár-használat Magyarországon (ohm) – egy kérdőíves szótárhasználati felmérés eredményei I. *Alkalmazott Nyelvtudomány*, 16(2):225–250. A letöltés ideje: 2021. 03. 10.

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P. és Vadász, N. (2019). `emtsv` — egy formátum mind felett. In Berend, G., Gosztolya, G. és Vincze, V. (szerk.): *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, 235–247, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Kumano, A. és Hirakawa, H. (1994). Building an MT dictionary from parallel texts based on linguistic and statistical information. In *The 15th International Conference on Computational Linguistics. Volume 1.*, 76–81, Kyoto, Japán. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

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

Maticsák, S. és Laihonen, P. (2011). Milestones in the History of Hungarian/Finnish Bilingual Lexicography. In Fábian, Z. (szerk.): *Hungarian lexicography I. Bilingual dictionaries*. Akadémiai Kiadó, Budapest.

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G. és Váradi, T. (2008). Methods and results of the hungarian wordnet project. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C. és Vossen, P. (szerk.): *Proceedings of GWC 2008*, 311–320, Szeged. University of Szeged, Department of Informatics.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. és Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. és Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*,

1659–1666, Kyoto, Japan. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Pirinen, T. A. (2015). 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(1):381–393.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Szinnyei, J. (1884). *Finn–magyar szótár*. Magyar Tudományos Akadémia, Budapest.

Tiedemann, J. és Nygaard, L. (2004). The OPUS corpus - Parallel Free. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. és Silva, R. (szerk.): *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 1183–1186, Lisbon. European Language Resources Association.

Wu, D. és Xia, X. (1994). Learning an English-Chinese Lexicon from a Parallel Corpus. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. és Silva, R. (szerk.): *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 206–213, Columbia, USA. Association for Machine Translation in the Americas.