Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban

Héja Enikő[1], Sass Bálint[1]

[1] MTA Nyelvtudományi Intézet
{eheja,sass.balint}@nytud.hu

Kivonat: Jelen cikk célja annak vizsgálata, hogy a párhuzamos korpuszokból fordítási ekvivalensek kinyerésére használt módszer kiterjeszthető-e többszavas kifejezésekre is. A kísérletben a többszavas kifejezéseket kizárólag igei szerkezetek alkották. Első lépésként a párhuzamos korpusz forrásnyelvi és célnyelvi oldalából külön-külön nyertük ki az igei szerkezeteket. A jelenlegi fázisban a kinyerés félig automatikus módon történt: előre meghatározott forrásnyelvi igékhez és ezek célnyelvi fordításaihoz tartozó igei szerkezeteket kerestünk, melyekből kézzel válogattuk ki a céljainknak megfelelőeket. A következő lépésben ezeket egytagú kifejezésekké vontuk össze a párhuzamos korpuszban. Az összevont igei szerkezetek már az illesztési algoritmus bemeneteként szolgálhattak. Eredményeink azt mutatják, hogy az alkalmazott módszer jól használható igei szerkezetek fordítási ekvivalenseinek detekciójára.

1 Bevezetés

Jelen cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX[1] projekt része. A projekt azt vizsgálja, hogy a nyelvtechnológiai módszerek és eszközök – különös tekintettel a párhuzamos korpuszokra – mennyiben járulhatnak hozzá a szótárkészítési folyamathoz. A szótárkészítés automatikus támogatása elsősorban a kevéssé használt nyelvek esetében bír jelentőséggel, hiszen az ilyen nyelvpárokra készült szótárak iránti kereslet alacsony, így a szükséges munkálatok finanszírozása is korlátozott. A projekt célkitűzése középméretű (kb. 15,000 szócikk), általános célú szótárak létrehozása volt a magyar-litván, illetve a francia-holland nyelvpárokra.

Jelenleg nem létezik olyan módszer, amely lehetővé tenné a szótárak *teljesen* automatikus előállítását. Ezért egy megfelelő lefedettségű és pontosságú lexikai erőforrás előállítása mindenképpen igényel emberi utószerkesztési munkálatokat is. Ennek fényében úgy fogalmazhatjuk meg feladatunkat, hogy célja a lexikográfusok számára olyan erőforrásokat biztosítani, amelyek a lehető legjobban csökkentik a teljes értékű, emberi felhasználásra alkalmas szótárak elkészítéséhez szükséges munkát. A fenti elvárásoknak megfelelő automatikusan generált erőforrásokat protoszótáraknak fogjuk nevezni a cikk hátralevő részében. Jelen írás az alábbi szerkezetet követi: a bevezetés után röviden ismertetjük az egytagú fordítási ekvivalensek kinyerésére használt

[1] http://www.efnil.org/projects/efnilex

módszert, illetve ennek előnyeit és hátrányait (2). Ezt követően vázoljuk a munkafolyamatot (3), amely két fő lépésre bontható: az igei szerkezetek kinyerésére (3.1), valamint a protoszótár létrehozására (3.2). Majd eredményeinket mutatjuk be (4), végül pedig a konklúziókat és a további teendőket (5).

2 Az alkalmazott módszer – előnyök és hátrányok

A statisztikai gépi fordítás térhódításával jelentősen megnőtt a párhuzamos korpuszok szerepe a nyelvtechnológiában. Már legalább 16 éve használnak különféle statisztikai algoritmusokat forrásnyelvi és célnyelvi szópárok kinyerésére, hogy így bővítsék a gépi fordítás bemenetéül szolgáló szótárakat (pl. [8]).

Érdekes módon a lexikográfusok között a mai napig sem eldöntött kérdés, hogy érdemes-e párhuzamos korpuszokra támaszkodni az emberi felhasználásra készülő szótárak készítése során. Például [1] szerint „*Túl magas az az ár, amit akkor kéne fizetni, ha a szótárszerkesztők kétnyelvű korpuszokat használnának. [...] Az ellenérvek: túl sok fordítási ekvivalens, a szerkesztés során mindegyik fordítási ekvivalens egyformán fontosnak tűnik a lexikográfus számára, [...], a bejegyzések a legtöbb felhasználó számára túl sok részletet tartalmaznak.*"

Eddigi kísérleteink [4] azt mutatták, hogy ha előfeldolgozásként szóillesztést végzünk, akkor a szóillesztés során kapott szópárok és ezek fordítási valószínűségei már megfelelő kiindulópontként szolgálhatnak egy kétnyelvű szótárhoz. Így tehát a fenti ellenvetések nem állják meg a helyüket. Először is, a forrásnyelvi szó és célnyelvi ekvivalense közötti fordítási valószínűség, valamint a forrásnyelvi szó és célnyelvi megfelelőjének gyakoriságai alapján szűrhetjük a fordítási jelölteket, így csökkentve a lehetséges fordítási ekvivalensek számát. Továbbá, a fordítási valószínűségek alapján sorrendezhetjük a fordítási jelölteket, ami biztosítja, hogy a leggyakrabban használt fordítás szerepeljen a szótári bejegyzés elején. Az általunk javasolt módszer további előnyei közé tartozik, hogy egy megfelelő méretű párhuzamos korpusz garantálja, hogy a legfontosabb fordítási ekvivalensek szerepelni fognak a szótárban. Ezenfelül, a párhuzamos korpusz gazdag tárháza a valódi nyelvből vett példamondatoknak, amelyek alapján a lexikográfus vagy a szótárhasználó kiválaszthatja a számára legmegfelelőbb fordítást a lehetséges fordítások közül. A fenti jellemzők miatt a javasolt módszer különösen alkalmas aktív[2] szótárak készítésének támogatására.

A [4]-ben javasolt módszer hátránya, hogy nem kezeli a többszavas kifejezéseket, így jelen állapotában alkalmatlan a több szóból álló fordítási ekvivalensek kiszűrésére. Ennek a feladatnak a megoldását mind a forrásnyelvi, mind a célnyelvi oldalon kiemelten fontosnak tartjuk, hiszen a többszavas kifejezések és szokásos fordításainak kinyerésével biztosíthatjuk, hogy a gyakori szófordulatok szerepeljenek a szótárban. Ez teszi lehetővé, hogy a szótár alapján természetesen hangzó célnyelvi szöveget hozhassunk létre. Jelen cikkben a többszavas kifejezések kezelését célzó első kísérletet ismertetjük. A cikkben ismertetett kísérlet célja *ige + bővítmény* szerkezetek fordítási ekvivalenseinek automatikus felismerése a francia-holland nyelvpárra. Reménye-

[2] Az aktív szótárak célja, hogy a forrásnyelvi anyanyelvi beszélőt segítsék a célnyelvi megnyilatkozások létrehozásában.

ink szerint az általunk javasolt módszer az ige+bővítmény szerkezetek fordítási ekvivalenseinek automatikus meghatározásával elősegíti a szótári tételek mikrostruktútájának kialakítását.

3 Munkafolyamat

A munkafolyamat két fő szakaszból áll. Az első lépésben a francia és holland igei szerkezetek félig automatikus kinyerésével hozzuk létre a vizsgálandó igei szerkezetek listáját (3.1). A második lépésben a kiválasztott többszavas igei szerkezeteket összevonjuk, így ezek az illesztés bemenetül szolgálhatnak. Eredményül egy többszavas igei szerkezeteket tartalmazó protoszótárat kapunk (3.2).

3.1 Francia és holland igei szerkezetek félig automatikus kinyerése

A kísérlethez a TLT-Centrale által fejlesztett Holland Párhuzamos Korpusz (DPC – Dutch Parallel Corpus) francia-holland alkorpuszát használtuk [5]. Az összesen 6,820,547 tokenes párhuzamos korpusz 186,945 illesztett egységet tartalmaz[3].

Első lépésben kézzel kiválasztottunk 20 gyakori, általunk poliszémnek gondolt francia igét (pl.: *mettre* 'tesz, helyez'). Ezek mindegyikéhez hozzárendeltünk egy-egy alapértelmezettnek tűnő holland fordítást (a *mettre* esetében ez a *leggen*). Az 1. táblázatban tételesen felsoroljuk a kiválasztott francia igéket, ezek holland megfelelőit, illetve magyar fordításukat. Megadtuk továbbá, hogy a következő lépés eredményeként hányféle különböző igei szerkezetet találtunk az adott igére.

1. táblázat: Francia igék és holland fordításaik.

Francia ige	Különböző igei szerkezetek	Holland fordítás	Különböző igei szerkezetek	Magyar fordítás
donner	12	*geven*	31	adni
effectuer	3	*teweegbrengen*	0	előidéz, véghezvisz
enlever	0	*verwijderen*	1	eltávolít
faire	31	*doen*	12	csinálni
mener	2	*leiden*	4	vezet
mettre	26	*leggen*	5	tenni

[3] Mivel a párhuzamos korpusz tartalmaz egy-a-többhöz, illetve több-az-egyhez megfeleltetéseket, a mondatok száma helyett a korpusz méretét az illesztett egységek számával adjuk meg.

montrer	4	*wijzen*	1	mutatni
obtenir	5	*behalen*	2	megszerezni
offrir	1	*aanbieden*	2	kínálni
ouvrir	1	*openen*	1	nyitni
passer	3	*vergaan*	0	eltölteni
porter	3	*brengen*	14	hozni
prendre	23	*nemen*	23	(el)venni
recevoir	2	*krijgen*	12	kapni
rendre	3	*maken*	19	tesz (vmilyenné)
rester	0	*blijven*	1	maradni
tenir	4	*houden*	11	tartani
traiter	1	*behandelen*	2	bánni
trouver	3	*vinden*	6	találni
voir	0	*zien*	0	látni

Amint azt a táblázat is mutatja, a kiválasztott igék némelyikéhez (*unlever*, *rester*, *voir, vergaan, zien*) egyetlen feltételeinknek megfelelő igei szerkezetet sem találtunk.

A következő lépésben automatikusan kinyertük a releváns francia, illetve holland ige+bővítmény szerkezeteket a párhuzamos korpusz megfelelő egynyelvű részéből. A [7]-ben leírt igei szerkezetek kinyerésére szolgáló módszert alkalmaztuk. Ez a módszer tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszon dolgozik. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzés pedig meg kell hogy állapítsa a tagmondat igéjét, a bővítmények fejét, valamint a bővítmények igéhez való szintaktikai viszonyát. A szintaktikai viszonyt a megfelelő esetrag vagy egy elöljárószó jelöli.

A módszer a tagmondatokban az ige mellett meglévő jellegzetes bővítménykereteket határozza meg, a gyakori részkeretek rendszerezett összeszámlálása révén. Előnye abban rejlik, hogy automatikusan felismeri, hogy melyik bővítménynél lényegi elem a konkrét fej és melyiknél csak az ige-bővítmény viszony, azaz egyszerre képes meghatározni az összetett igéket és a vonzatkereteket is. A *hasznot húz vmiből* szerkezet esetén például felfedezi, hogy a lexikálisan kötött tárgy mellett egy *-ból*/*-ből* esetragos vonzat szerepel az igei keretben.

Az algoritmus vázlata a következő. Vesszük a korpusz összes tagmondatát. Előállítjuk a tagmondatoknak megfelelő szerkezeteket, melyekben a bővítményi fejeket minden variációban, váltakozva töröljük, illetve megtartjuk. Hossz szerint csökkenő sorba rendezzük a kapott szerkezetlistát, majd sorra elhagyjuk azokat a szerkezeteket, melyeknek a gyakorisága 5-nél kisebb, és ezek gyakoriságát a megfelelő, illeszkedő rövidebb keret gyakoriságához adjuk. A megmaradó szerkezetek gyakoriság szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

Említettük, hogy az algoritmus bemenetként tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszt vár. Jelen kísérletben mindkét elemző lépést egyszerű szabályok használatával közelítettük a francia és a holland nyelv esetében is. Tagmondathatárt jelent nyilvánvalóan a mondathatár. Ezen kívül a kötőszó, az alárendelt tagmondatot bevezető holland *te*, ill. francia *pour*, a vonatkozó névmás és bizonyos írásjelek (vessző, kettőspont és pontosvessző) is, amennyiben a legutóbbi tagmondathatár óta szerepelt a mondatban ige. Bővítményi fejek a főnevek (valamint a reflexív igék miatt a holland *zich* és a francia *se*); a bővítményi viszonyt pedig a szó előtt álló prepozíció jelzi. Prepozíció híján az ige előtt alanynak, az ige után pedig tárgynak vettük az adott bővítményt.

Eredményként olyan összetett igei szerkezeteket kapunk, mint a francia *mettre accent sur*... vagy holland megfelelője, *a leggen nadruk op*... (magyarul mindkettő: 'hangsúlyt helyez vmire').

Az így kinyert kifejezésekből kézzel válogattuk ki azokat az igei szerkezeteket, amelyekről azt gondoltuk, hogy nem, vagy csak részben kompozicionálisak. Mivel fordítási feladatról van szó, a kompozicionalitás ebben az esetben nem önmagában, hanem egy másik nyelv függvényében értelmezhető. Így a nem transzparens kifejezések mellett intézményesült kifejezések kinyerésére is törekedtünk. Például bár a *mettre l'appareil hors tension* (szó szerint 'feszültségen kívül helyezni a készüléket' – 'áramtalanítani') francia kifejezés kompozicionális szerkezetnek tekinthető, bekerült a listánkba, mivel a holland fordítás – *uitschakelen* – már nem őrzi meg a kifejezés eredeti szerkezetét. Így tehát az automatikusan kinyert igei szerkezetek közül azokat vettük fel a listánkba, amely az alábbi kritériumok közül bármelyiknek megfelelt:

(1) A kifejezés jelentése az eredeti nyelven nem transzparens (pl. *faire mouche* 'célt érni')
(2) Ha feltételezhető, hogy az igei szerkezet fordítása nem tükörfordítás
    a. Az igei szerkezet intézményesült
    b. Az igei szerkezet magyar fordításában a főnév igemódosítóként[4] jelenik meg (l. [3]).

Fontos hangsúlyozni, hogy az igei minták közül a kollokációszerű mintákat választottuk ki, azaz azokat, melyek az ige mellett tartalmaztak egy konkrét főnevet is. A konkrét főnév az alanyeset kivételével bármilyen esetben állhat. Az igei szerkezetek kiválasztásakor nem törekedtünk a teljes igei vonzatkeret megőrzésére, így bizonyos esetekben a kitöltetlen – vagyis tipikus főnévi lemma nélkül álló – esetragokat el-

[4] Igemódosítónak azokat a névelőtlen névszókat tekintettük, amelyek az igekötővel nagyjából hasonló disztribúciót mutatnak és semleges, tagadást nem tartalmazó mondatban közvetlenül az ige előtt állnak.

hagytuk. Ennek oka egyfelől, hogy az igei szerkezetek összevonásával csökkenthetjük az adathiány problémáját, másfelől pedig az, hogy mivel az illesztés bemeneti korpusza nem tartalmazott sem részleges szintaktikai elemzést, sem tagmondatfelismerést, az esetek egy jelentős részében lehetetlen volt pontosan azonosítani a megfelelő prepozíciót.

Az illesztés bemenetéül az alábbi szintaktikai mintákra illeszkedő igei szerkezetek szolgáltak (V: ige; N_ACC: a főnévi lemma szintaktikai funkciója tárgy; ACC: kitöltetlen tárgy; N_PREP: a főnévi lemma valamilyen prepozícióval szerepel; PREP: kitöltetlen prepozíció):

(1) V + N_ACC
(2) V + N_PREP
(3) V + ACC + N_PREP
(4) V + N_ACC + PREP
(5) V + N_PREP + PREP
(6) V + N_PREP + N_PREP + PREP

A harmadik lépésben következik ezen igei szerkezetek korpuszbeli azonosítása, összevonása és illesztése.

3.2 A protoszótár létrehozása

A továbbiakban a kiválogatott többszavas igei kifejezéseket egy szóként kezeltük, és így közvetlenül alkalmaztuk az eredeti szavakon működő illesztő algoritmust.

Mint említettük, az illesztés bemeneti korpusza lemmatizált, de nem tartalmaz sem tagmondat-információt, sem részleges szintaktikai elemzést. Ezen munkaszakasz első lépése a minták felismerése a korpuszban, majd ezek összevonása egytagú kifejezéssé. A 126 francia igei szerkezet összesen 7805-ször, míg a 146 holland igei szerkezet 8029-szer fordult elő a párhuzamos korpuszban.

A szóillesztést a GIZA++ szoftverrel végeztük [6], amely a szóillesztés során fordításjelölteket hoz létre, úgy, hogy a forrásnyelvi és célnyelvi lemmapárokhoz fordítási valószínűséget rendel. A fordítási valószínűség a célnyelvi és forrásnyelvi szópár feltételes valószínűségének közelítése – $P(\text{szó}_{\text{cél}}|\text{szó}_{\text{forrás}})$ – az EM (expectation maximization) algoritmus alapján [2].

A protoszótárak kiindulási alapját az így kinyert fordítási jelöltek és fordítási valószínűségeik képezték. Mivel a fordítási valószínűség 0-tól 1-ig bármilyen értéket felvehet, ebben a szakaszban még sok helytelen fordítási jelöltünk van. Ezért szükség van olyan szűrők bevezetésére, amelyek lehetővé teszik a legjobb fordításjelöltek automatikus kiválasztását a lehető legtöbb helyes fordításjelölt megtartásával. Eddigi tapasztalataink azt mutatták [4], hogy a fordítási valószínűségek és a forrásnyelvi, illetve célnyelvi korpuszgyakorisági adatok együttesen már jól használhatóak az eredmények szűrésére. Így a protoszótárban az alábbi adatok szerepelnek:

2. táblázat: A protoszótárban szereplő adatok.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
mettre_à_jour 'frissít'	*actualiseren* 'frissít'	0.0472766	105	39

A már elvégzett kiértékelések alapján (magyar-litván, magyar-szlovén) az alábbi általános feltételeket fogalmazhatjuk meg a protoszótárban szereplő tételekkel szemben:

(1) A forrásnyelvi és a célnyelvi szónak is legalább 5-ször elő kell fordulnia a párhuzamos korpuszban. Ez a feltétel szükséges ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez.

(2) Mivel a fordítási valószínűség a célnyelvi szó feltételes valószínűsége a forrásnyelvi szó mellett, további követelmény, hogy a két lemma (vagy kifejezés) gyakorisága ne legyen túl különböző. Gyakori célnyelvi lemmák esetében az illesztési algoritmus magas fordítási valószínűséget rendelhet rossz fordítási jelöltekhez is, ha a forrásnyelvi lemma ritkán fordul elő a korpuszban és a lemmák gyakran fordulnak elő illesztett mondatokban. A 3. táblázatban egy ilyen fordítási jelöltpárt mutatunk be.

3. táblázat: Rossz fordítási jelöltpár.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
mettre_vie_en_danger 'veszélyezteti az életét'	*rekening_houden* 'figyelembe vesz'	0.876763	24	577

Mivel a munka célja emberi felhasználásra szánt szótárak automatikus előállítása, a kiértékelés során a rossz és a jó fordítási jelöltpárok helyett *lexikográfiailag hasznos* és *lexikográfiailag nem hasznos* fordítási jelölteket különböztettünk meg. Az egyszavas kifejezések kiértékelése azt mutatta, hogy a vizsgált paraméterek mellett (a forrásnyelvi szó és a célnyelvi szó gyakorisága legalább 5 és a fordítási valószínűség legalább 0.5) a találatok mintegy 90%-a lexikográfiailag hasznos fordítás. Tekintve, hogy a pontosság fordítottan korrelál a lefedettséggel, és utólagos kézi feldolgozásra mindenképpen szükség van, a paraméterek olyan beállítására lesz szükség, amely kb. 60%-os pontosságot eredményez. Egy további megfigyelés szerint a gyakoribb lemmák esetén már a jóval kisebb fordítási valószínűséggel rendelkező párok jelentős része is a lexikográfiailag hasznos kategóriába tartozik, így ezek figyelembevétele tovább növeli a lefedettséget. Mivel a jelen munkaszakasz feladata nem a megfelelő szűrési paraméterek beállítása, hanem annak vizsgálata, hogy az eredeti módszer kiterjeszthető-e többszavas kifejezésekre is, a paramétereket úgy állítottuk be, hogy a lehető legtöbb jó fordítási jelöltet megtartsuk, még akkor is, ha ez alacsony pontossághoz vezet. Ezért minden olyan fordítási jelöltpárt megtartottunk, ahol a pár mindkét tagjának gyakorisága legalább 5, a fordítási valószínűséget azonban 0.02-re csökkentettük. A következő szakaszban eredményeinket mutatjuk be.

4 Eredmények

Az eredmények részleges kiértékelése azt mutatta, hogy a használt módszer alkalmas igei szerkezetek fordításainak kinyerésére.

A fenti paraméterekkel összesen 906 olyan fordítási jelöltet kaptunk, amelynek legalább egyik tagja többszavas kifejezés, ebből 632 esetben a forrásnyelvi kifejezés többszavas. Ez 113 különböző francia igei szerkezetet jelent (a fordításjelöltek között 127 különböző holland igei szerkezet szerepel). A részleges kiértékelés során a francia többszavas kifejezésekre koncentráltunk.

294 fordítási jelöltpárt értékeltünk ki kézzel. A kiértékelés során a *teljesen jó* és a *rossz fordítási* párok mellett megkülönböztettünk *részlegesen jó* fordítási párokat is. A részlegesen jó fordítási párok esetében a jelöltpár valamelyik tagja egy fel nem ismert többszavas kifejezés, így csak részleges illesztés történt. A kiértékelt párok közül 57 fordítás volt tökéletes (19%) és 28 fordítási jelölt bizonyult részlegesen jó fordításnak. A választott paraméterek mellett 189 fordítási jelölt volt teljesen rossz. Mivel ebből 132 csak egy mondatpárban fordult elő a teljes párhuzamos korpuszban, a jövőben egy további paraméterként azon mondatpárok számát is felvesszük, amelyben a fordítási jelöltek előfordulnak. Ezen szűrő hasznosságát támasztja alá, hogy a 149 fordítási jelölt közül, amelyek csak egy mondatpárban fordultak elő, csak 17 volt lexikográfiailag hasznos fordítás.

Az eredmények kézi ellenőrzése során azt találtuk, hogy a rossz fordítási jelöltek száma jelentősen csökkenthető lenne, ha a részleges szintaktikai elemzésre és a tagmondathatárra vonatkozó információkat az illesztés során is figyelembe vennénk.

Fontos hangsúlyozni, hogy jelen cikk célja nem a lehető legnagyobb pontosság elérése, hanem a módszer alkalmazhatóságának vizsgálata igei szerkezetek kinyerésére. A paraméterek átállításával a pontosság jelentősen növelhető. A 4. és 5. táblázatban két példával illusztráljuk, hogy a megközelítés hogyan használható fordítások kinyerésére.

4. táblázat: Első példa.

$\text{Kifejezés}_{\text{forrás}}$	$\text{Kifejezés}_{\text{cél}}$	$P(\text{szó}_{\text{cél}}\|\text{szó}_{\text{forrás}})$	$\text{Gyak}_{\text{forrás}}$	$\text{Gyak}_{\text{cél}}$
mettre_à_jour	*bijwerken*	0.0649992	105	60
FR: Comment les **met**-on **à jour**?				
NL: Hoe worden ze **bijgewerkt**?				
H: Hogyan lehet ezeket **frissíteni**?				
mettre_à_jour	*actualiseren*	0.0472766	105	39
FR: De plus, un PGR **mis à jour** doit être soumis:				
NL: Bovendien dient een **geactualiseerd** RMP ingediend te worden :				

H: Ezenfelül, egy **frisített** PGR-t kell elküldeni:				
mettre_à_jour	*aanpassing*	0.0372093	105	442
FR: **Mise à jour** de la liste des produits admis au remboursement				
NL: **Aanpassing** van de lijst van de voor vergoeding aangenomen producten				
H: A költség-visszatérítésre elfogadott termékek listájának **kiigazítása**				
mettre_à_jour	*update*	0.029671	105	34
FR: Toutes les informations au sujet du changement y ont été publiés avec de fréquentes **mises à jour** .				
NL: Alle informatie met betrekking tot de omslag is erop gepubliceerd , met regelmatige **updates**.				
H: Minden változásra vonatkozó információ ott van közzétéve, rendszeres **frissítésekkel**.				

5. táblázat: Második példa.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
prendre_en_considération	*nemen_in_aanmerking*	0.186464	93	73
FR: Les offres qui dérogent à cette date ne sont pas **prises en considération**.				
NL: Offertes die hiervan afwijken worden niet **in aanmerking genomen**.				
H: A megadott dátumoktól eltérő ajánlatokat nem **vesszük figyelembe**.				
prendre_en_considération	*houden_rekening*	0.166621	93	438
FR: Concernant votre apport financier personnel , vos revenus sont **pris en considération**.				
NL: Voor uw eventuele persoonlijke financiële bijdrage wordt **rekening gehouden met** uw inkomsten.				
H: Az Ön személyes anyagi hozzájárulásánál **figyelembe vesszük** a jövedelmét.				
prendre_en_considération	*nemen_in_overweging*	0.0221903	93	35
FR: La date de conclusion à **prendre en considération** pour le choix …				
NL: De datum van sluiting die **in overweging** moet worden **genomen** voor de keuze …				
H: A zárás időpontja, amelyet a választáshoz **figyelembe** kell **venni** ….				

5 Konklúziók és további teendők

Jelen cikk célja annak vizsgálata volt, hogy a párhuzamos korpuszokból fordítási ekvivalensek kinyerésére használt módszer kiterjeszthető-e többszavas kifejezésekre is. A kísérletben a többszavas kifejezéseket kizárólag igei szerkezetek alkották. Első lépésként a párhuzamos korpusz célnyelvi és forrásnyelvi oldalából külön-külön nyertük ki az igei szerkezeteket. A jelenlegi fázisban a kinyerés félig automatikus módon történt: az előre meghatározott forrásnyelvi igékhez és ezek célnyelvi fordításaihoz tartozó igei szerkezeteket kerestünk, melyekből kézzel válogattuk ki a céljainknak megfelelő igei szerkezeteket. A következő lépésben ezen igei szerkezeteket egytagú kifejezésekké vontuk össze a korpuszban. Így az összevont igei szerkezetek az illesztési algoritmus bemeneteként szolgálhattak. Annak ellenére, hogy a kiértékelés során a szűrésre használt paramétereknek alacsony értékeket adtunk meg, és így eredményeink meglehetősen alacsony pontosságúak, a kinyert igei szerkezetek egyértelműen mutatják, hogy a módszer jól használható igei szerkezetek fordítási ekvivalenseinek detekciójára.

További feladataink közé tartozik a részleges szintaktikai elemzés és a tagmondathatár-felismerés minőségének javítása a forrásnyelvre és a célnyelvre egyaránt, valamint ezen információ figyelembevétele a szóillesztés bemeneti korpuszában is.

A lefedettség jelentősen növelhető lenne, ha nemcsak előre kiválasztott igékhez tartozó igei szerkezeteket illesztenénk, hanem minden igei szerkezetet, amely gyakran fordul elő a párhuzamos korpuszban. Célunk továbbá, hogy a célnyelvi igei szerkezetek detekciójánál ne a forrásnyelvi igék feltételezett fordításából induljunk ki, hanem ezeket egymástól függetlenül nyerjük ki. Elvárásaink szerint egy ilyen lépés csökkentené a részleges illesztések számát is, és növelné a teljesen jó fordításai ekvivalensek arányát.

További tervezett kutatási irány a módszer kiterjesztése a kollokációkra is.

Bibliográfia

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
2. Dempster, A. P., Laird, N. M., Rubin. D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B Vol. 39 No. 1 (1977) 1–22
3. É. Kiss, K.: Mondattan. In: É. Kiss, K., Kiefer, F., Siptár, P. (szerk.): Új magyar nyelvtan. Osiris Kiadó, Budapest (2003) 15–184
4. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography.In: Proceedings of the LREC2010 Conference. La Valletta, Malta (2010) 2798–2805
5. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007. Birmingham, United Kingdom (2007)
6. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics Vol. 29 No. 1 (2003) 19–51

7. Sass, B.: A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In: Angelova G. et al. (szerk.): Proceedings of RANLP 2009. Borovec, Bulgária (2009) 399–403
8. Wu, D.: Learning an English-Chinese Lexicon from a Parallel Corpus. In: Proceedings of AMTA'94 (1994) 206–213