Igei bővítménykeretek fordítási ekvivalenseinek kinyerése mélyen elemzett párhuzamos korpuszból

Héja Enikő[1], Takács Dávid[1], Sass Bálint[1]

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

Kivonat: Jelen cikk célja annak vizsgálata, hogy a mély szintaktikai elemzés növeli-e a fedést és a pontosságot igei szerkezetek fordítási megfelelőinek teljesen automatikus kinyerése során. Első lépésként a párhuzamos korpusz forrásnyelvi és célnyelvi oldalát külön-külön elemeztük, majd ebből nyertük ki az igei szerkezeteket egy felügyelet nélküli tanuló algoritmussal. Az így előállt igeiszerkezet-listát gyakorisági alapon szűrtük. A következő lépésben az igei szerkezeteket egytagú kifejezésekké vontuk össze a párhuzamos korpuszban, hogy az egytokenes igei szerkezetek az illesztési algoritmus bemeneteként szolgálhassanak. 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 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 (min. 15,000 szócikk), általános célú szótárak létrehozása volt a magyar-litván, illetve a francia-holland nyelvpárokra.

A statisztikai gépi fordítás térhódításával jelentősen megnőtt a párhuzamos korpuszok szerepe a nyelvtechnológiában. Érdekes módon a lexikográfusok között nem eldöntött kérdés, hogy használhatóak-e a párhuzamos korpuszok emberi felhasználásra készülő szótárak előállítására (l. pl. [1]). Eddigi kisérleteink azt mutatták, hogy ha előfeldolgozásként szóillesztést végzünk, akkor az általunk javasolt módszer számos előnnyel rendelkezik a hagyományos lexikográfiai módszertannal szemben [5]. A javasolt módszer hátránya, hogy nem kezeli a többszavas kifejezéseket, így önmagában alkalmatlan a több szóból álló fordítási ekvivalensek kiszűrésére. Ennek a feladatnak a megoldása kiemelten fontos, hiszen egy szótárnak tartalmaznia kell azokat a többszavas kifejezéseket is, amelyek fordítása nem kompozicionális.

[6], illetve [9] alátámasztották, hogy egy előfeldolgozó modul hozzáadása elvileg lehetővé teszi a többszavas *ige* + *bővítmény* szerkezetek fordítási megfelelőinek automatikus kinyerését. Eredményként olyan összetett igei szerkezeteket kapunk, mint a

francia *faire partie de...* vagy holland megfelelője, a *deel uitmaken van...* (*részét képezi vminek*).

Feladatunk a módszert továbbfejleszteni úgy, hogy a kinyert párhuzamos igei szerkezetek felvehetőek legyenek a szótárba: vagyis a pontosság és a fedés növelésére egyaránt szükség van. Ennek érdekében a kutatás jelen szakaszában a [6]-ban, illetve [9]-ben leírtakat az alábbiak szerint módosítottuk. (1) Előre meghatározott igék helyett minden elegendően gyakori igét figyelembe vettünk, (2) minden igei szerkezet a vizsgálat tárgyát képezi, nemcsak azok a szerkezetek, amelyek főnévi lemmát is tartalmaznak, (3) részlegesen elemzett párhuzamos korpusz helyett mély szintaktikai annotációval rendelkező párhuzamos korpuszt használtunk az igei szerkezetek kinyeréséhez.

Azt várjuk, hogy a javasolt módszer az ige+bővítmény szerkezetek fordítási ekvivalenseinek teljesen automatikus meghatározásával hozzájárul a szótári tételek mikrostruktútájának kialakításához.

A következő szakaszban vázoljuk a munkafolyamatot (2), amely három fő lépésből áll: a párhuzamos korpusz szintaktikai elemzése (2.1), az igei szerkezetek automatikus kinyerése (2.2), valamint a protoszótár létrehozása (2.3). Majd eredményeinket mutatjuk be (3), végül pedig a konklúziókat és a további teendőket (4).

2 A munkafolyamat

A munkafolyamat három fő szakaszból áll. Az első lépésben elvégezzük a párhuzamos korpusz francia és holland részének mély szintaktikai elemzését, majd az így előállt frázisstruktúra-szerkezeteket az igei szerkezet kinyerő algoritmus által megkövetelt részleges függőségi elemzésekké konvertáljuk (2.1). A második lépésben a francia és holland igei szerkezetek egymástól független automatikus kinyerésével létrehozzuk a vizsgálandó igei szerkezetek listáját (2.2). A harmadik lépésben a kiválasztott többszavas igei szerkezeteket egytokenes kifejezésekké vonjuk össze, így ezek az illesztés bemenetül szolgálhatnak. Eredményül egy többszavas igei szerkezeteket tartalmazó protoszótárat kapunk (2.3).

2.1 A holland-francia párhuzamos korpusz szintaktikai elemzése

A kísérlethez a TLT-Centrale által fejlesztett Holland Párhuzamos Korpusz (DPC – Dutch Parallel Corpus) francia-holland alkorpuszát használtuk [7]. Az összesen 6,820,547 tokenes párhuzamos korpusz 186,945 illesztett egységet tartalmaz.

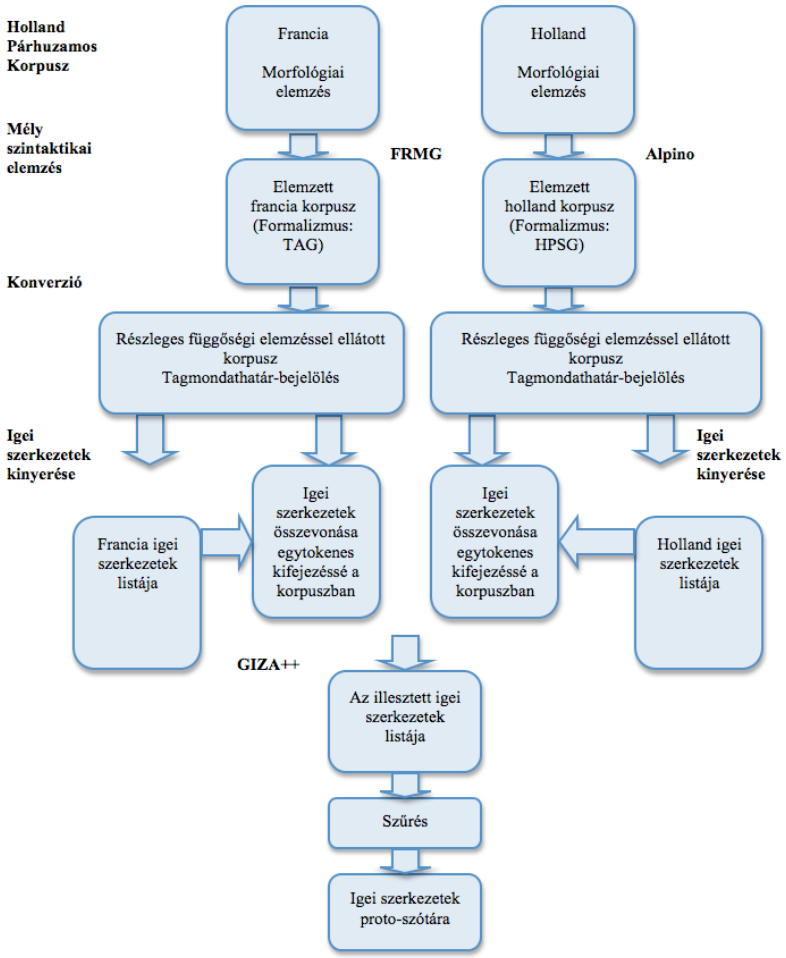
A holland esetében a HPSG elemzést végző Alpinot [2] használtuk, míg a francia korpuszt az FRMG hibrid TIG/TAG-parszerrel elemeztük[1] [11].

Az Alpino szabályalapú szintaktikai elemző a párhuzamos korpusz holland részkorpuszát részletetes annotációval látja el: megjelöli a frázisok határait és megadja a frázisok szintaktikai funkcióit. Ennek során felismeri az igéhez tartozó vonzatokat és partikulákat. Elvégzi a frázisok belső elemzését is: annotációval látja el a frázis fejét

[1] A szövegek elemzéséért köszönettel tartozunk Gábor Katának.

és a fejhez tartozó dependenseket. Az Alpino számunkra kiemelten fontos tulajdonsága, hogy felismeri a tagmondathatárokat, és megadja a tagmondatok egymáshoz való viszonyát (főmondat, mellékmondat, koordináció).

Az FRMG hasonló mélységű elemzést végez, mint az Alpino. Egy fontos különbség azonban, hogy az elemzés nem tartalmazott tagmondathatárra vonatkozó információt, ezért a tagmondathatár-felismerést saját szabályokkal végeztük el, amelyeket később részletezünk.



1. ábra: A munkafolyamat.

A következő lépésben az Alpino és az FRMG parszer kimenetét külön-külön részleges függőségi elemzéssé alakítottuk, hogy az elemzett korpuszok az igekinyerő algoritmus bemenetéül szolgálhassanak.

Az igei szerkezeteket kinyerő algoritmus abból az előfeltevésből indul ki, hogy (1) az ige jellemző bővítménykeretét mindig az a tagmondat tartalmazza, amelyben az ige előfordul, (2) egy tagmondat csak egy igéhez tartozó bővítményeket tartalmaz. Ebből következően a konverzió során meg kellett oldani a tagmondathatár-felismerést a francia esetében, valamint visszaállítani a teljes vagy eredeti bővítménykeretet azokban az esetekben, amikor erre szükség volt (pl. passzív igék, határozói és melléknévi igeneves szerkezetek). Ezeket utólagos átalakító szabályok hozzáadásával valósítottuk meg. A szabályok a részletes szintaktikai annotáción alapulnak, amely azt is jelöli, ha az ige valamilyen képzett formában szerepel (passzív, illetve különféle igeneves szerkezetek).

A holland esetében az alábbi átalakításokat végeztük el:

(1) Passzív szerkezetek aktívvá alakítása
(2) Segédigék törlése az összetett igeidők esetében
(3) Melléknévi igeneves szerkezetek konverziója tagmondattá

A francia elemzés esetében a fentieken túl a tagmondathatárok bejelölésére is szükség volt, így a fenti szabályokhoz továbbiakat adtunk hozzá:

(4) Melléknévi igeneves szerkezetek önálló tagmondatot alkotnak
(5) A vonatkozó névmások előtt legyen tagmondathatár
(6) A főnévi igenév előtt is van tagmondathatár, ha a főnévi igenév előtt valamilyen prepozíció áll (*de*, *pour*, *sans*, *en vue de*, *à* stb.)
(7) Legyen tagmondathatár koordinált tagmondatokat összekötő kötőszavak helyén (*et* - *és*, *puis* - *aztán*, *ou* - *vagy*, stb.)
(8) Legyen tagmondathatár az alárendelt mondatokat bevezető kötőszavak helyén (*que* - *hogy*, *quand*, *pendant que* - *amikor*, stb)
(9) Ha két ige között nincs tagmondathatár, akkor szúrjon be tagmondathatárt vessző, pontos vessző vagy kettőspont esetén.

Végül el kellett döntenünk, hogy a részletes szintaktikai annotáció mely jegyeit kívánjuk figyelembe venni az igei bővítménykeretek kinyeréséhez. Itt két ellentmondó követelménynek kell eleget tenni: egyfelől, minél több jegyet tartunk meg az eredeti elemzésből, annál részletesebben karakterizálhatjuk az igei bővítménykereteket. Másfelől, túl sok jegy alkalmazása jelentősen ronthatja az eredményeket, hiszen az irreleváns címkék növelik az adatok diverzitását. A típusok számának növekedésével párhuzamosan csökken a típusok előfordulási gyakorisága, ez pedig rontja a generált szótár minőségét.

Első megközelítésben megtartottuk az igét, az igével közvetlenül függőségi viszonyban levő összetevő fejét, valamint a fej dependensei közül az esetleges mellékneveket, illetve egyéb módosítókat a vonzatos főnevek esetében, míg a névelőket elhagytuk. A koordinált szerkezetekből (ha nem koordinált tagmondatokról volt szó) mindig csak az első összetevőt őriztük meg. A következő részben látni fogjuk, hogy bizonyos esetekben ez is túl részletes elemzésnek bizonyult, így további empirikus vizsgálatot igényel, hogy pontosan milyen mélységű elemzést érdemes végezni.

2.2 Az igei szerkezetek automatikus kinyerése

A releváns francia és holland *ige+bővítmény* szerkezeteket automatikusan nyertük ki a párhuzamos korpusz megfelelő egynyelvű részeiből. Az igei szerkezetek automatikus kinyerése során az ige mellett meglévő jellegzetes bővítménykereteket határozzuk meg a tagmondatokban a gyakori részkeretek rendszerezett összeszámlálása révén. A [9]-ben részletesen leírt módszer előnye abban rejlik, hogy felismeri, hogy melyik bővítménynél lényegi elem a konkrét fej és melyiknél csak az ige-bővítmény viszony. Így egyszerre képes meghatározni az összetett igéket és a vonzatkereteket is. A *haszнot húz* vmiből szerkezet esetén például felfedezi, hogy a lexikálisan kötött tárgy mellett egy *-ból*/*-ből* esetragos vonzat is 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.

Az igeiszerkezet-kinyerő módszer alapvetően 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 elöljáró-szó jelöli. Mivel az igei szerkezet fogalmát a vonzatkeretnél tágabban értjük, mély szintaktikai annotációval rendelkező korpuszokon is futtatható az algoritmus úgy, hogy többletinformációt nyerjünk ki belőle (az algoritmus az igei vonzatokon túl a jellemző bővítményeket is megadja – akkor is, ha azok szabad határozók – sőt az igei szerkezet részét képezik a jellemző lexikai fejek is). Az 1. és 2. táblázatban példákat láthatunk az automatikusan kinyert igei szerkezetekre.

1. táblázat. A holland *'gebruiken'* ige négy leggyakoribb szerkezete.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
gebruik obj1	470	*használ vmit*
gebruik niet=mod:ADV obj1	159	*nem használ vmit*
gebruik obj1 obj1_ADJ	104	*használ vmilyen vmit*
gebruik obj1 als=predc:CP	95	*úgy használ valamit, hogy ...*

Az 1. táblázat mutatja azt is, hogy a részletes elemzés eredményeképpen a *'nem használ vmit'* illetve a *'használ valamilyen vmit'* is gyakori kereteknek minősülnek, ám felvételük egy igei kereteket tartalmazó szótárba a keretek kompozicionalitása miatt nem indokolt. A megfelelő bővítmények elhagyásával mindkét keret a *'használ vmit'* kerethez sorolódna, így növelve ezen keret gyakoriságát a korpuszban, és ezáltal a megfelelő fordítási ekvivalensek kinyerésének a valószínűségét.

A 2. táblázatban szintén szerepelnek irreleváns keretek is a mély szintaktikai elemzés eredményeként:

2. táblázat. A holland *'geven'* ige négy leggyakoribb szerkezete.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
geef obj1	170	*ad vmit*
geef obj1 obj1_ADJ	80	*ad vmilyen vmit*
geef aan:obj2 obj1	78	*ad vkinek vmit* (indirekt)
geef obj1 obj2	72	*ad vkinek vmit* (direkt)

A táblázatban látszik, hogy ha a tárgyat módosító jelzőt nem vennénk figyelembe, akkor a *'geven'* leggyakoribb szerkezetei pontosan az „elvártak" lennének.

A 3. táblázatban található példa már lexikai bővítményt is tartalmaz a jellemző esetkeret mellett. Ez a mély elemzés egy másik nem kívánt hatását szemlélteti: a parszer ugyanahhoz a felszíni szerkezethez bizonyos esetekben különböző annotációkat rendel, és ez – függetlenül attól, hogy melyik a jó elemzés – megint csak a rendelkezésre álló adatok csökkenéséhez vezet.

3. táblázat: A holland *'een beroep doen op'* elemzései.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
doe beroep=obj1 obj1_op	72	*felhívást tenni vmire*
doe beroep=obj1 op:mod	39	*felhívást tenni vmire*

Az első esetben a holland *'op'* (-rA) az ige tárgyának, a *'beroep'*-nak, míg a második esetben magának az igének a bővítménye. További probléma, hogy ennek a szerkezetnek a névelő (*een*) kötelezően része, de ez mindkét keretből hiányzik.

A következő lépésben automatikusan választottuk ki azokat az igei szerkezeteket, amelyeket akár forrásnyelvi, akár célnyelvi oldalon a szótárban szerepeltetni akartunk. Egy lehetséges megközelítés, hogy heurisztikát dolgozunk ki a „*lexikográfiai szempontból érdekes*" bővítménykeretek automatikus szűrésére. Mivel fordítási feladatról van szó, a kompozicionalitás ebben az esetben nem önmagában, hanem egy másik nyelv függvényében értelmezhető. A javasolt módszer egyik kiemelten fontos tulajdonsága a nyelvfüggetlenség. Így elképzelhető, hogy *A* nyelv egy igei szerkezete kompozicionálisan fordul le *B* nyelvre, de nem kompozicionális *C* nyelven. Ebben az esetben tehát azt kell mondanunk, hogy *A* nyelv adott kifejezése lexikográfiailag érdekes az első esetben, és érdektelen a másodikban. A nyelvfüggetlenség miatt járhatóbb megközelítési módnak tűnik az igei szerkezeteket gyakorisági alapon szűrni. Ebben az esetben feltételezzük, hogy egy szótárban a gyakran előforduló jelenségeket célszerű rögzíteni, függetlenül attól, hogy ezek fordítása transzparens-e vagy sem egy másik nyelven.

Így tehát az automatikusan kinyert igei szerkezetek közül azokat vettük fel a listánkba, amelyek legalább ötször előfordultak a párhuzamos korpusz megfelelő oldalán. Ennek a kritériumnak a holland oldalon 289 ige felelt meg, összesen 5804 kerettel, míg a francia igelista 391 igét tartalmazott 5987 különböző kerettel.

2.3 A keretek azonosítása, összevonása és a protoszótár létrehozása

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

[6]-ban csak azokat a szerkezeteket vizsgáltuk, amelyek az igén kívül is tartalmaztak valamilyen kötött lexikai elemet. Az igei szerkezetek kiválasztásakor nem törekedtünk a teljes bővítménykeret megőrzésére, így bizonyos esetekben a kitöltetlen – vagyis tipikus főnévi lemma nélkül álló – esetragokat elhagytuk. Ennek oka egyfelől az volt, hogy az eltérő igei szerkezetek összevonásával növelhettük a szükséges adatok mennyiségét. Másfelől, 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.

Ezzel szemben a jelen kísérlet célja minden megfelelően gyakori igei bővítménykerethez fordítási megfelelőt találni, függetlenül attól, hogy tartalmaz-e kötött lexikai elemet. Az ige bővítményeit értelemszerűen csak az igéhez tartozó tagmondatban kerestük. Az illeszkedő igei keretek közül a leghosszabbakat választottuk, és ezt vontuk össze a párhuzamos korpusz elemzett változatában.

Míg az említett első kísérletben a 126 francia igei szerkezet összesen 7805-ször, és a 146 holland igei szerkezet 8029-szer fordult elő a párhuzamos korpuszban, addig a jelen kísérletben 170,229 illeszkedő francia bővítménykeret és 207,610 illeszkedő holland bővítménykeretet találtunk a párhuzamos korpuszban.

A továbbiakban a kiválogatott többszavas igei kifejezéseket egy tokenként kezeltük és így közvetlenül alkalmaztuk az működő illesztő algoritmust.

Az illesztést a GIZA++ szoftverrel végeztük [8], amely az illesztés 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 [3].

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 [5], 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:

4. táblázat. Francia és holland fordítási jelöltpárok és paramétereik.

Kifejezés$_{forrás}$	**Kifejezés$_{cél}$**	**P(szó$_{cél}$|szó$_{forrás}$)**	**Gyak$_f$**	**Gyak$_c$**
prendre médicament=obj1	neem_in genees_middel=obj1	0.377261	53	32
	gebruik genees_middel=obj1	0.102349	53	21
	start gebruik=met:cmp met:cmp_van	0.0971227	53	28
	sta onder_invloed=particle drug=van:cmp	0.050697	53	11

A 4. táblázatban látható, hogy a francia *'prendre médicament*' (*gyógyszert bevenni*) szerkezetnek a legvalószínűbb holland megfelelője az '*geneesmiddel innemen*'. Ezt követi a *'geneesmiddel gebruiken'* (*gyógyszert használni*). A *'start met gebruik van'* nem teljes keret (*elkezdeni a használatát valaminek*) szintén releváns fordításnak tekinthető. A legkevésbé valószínű, ám lexikográfiai szempontból még érdekes fordítás a '*staan onder invloed van drug*' (drog hatása alatt állni).

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

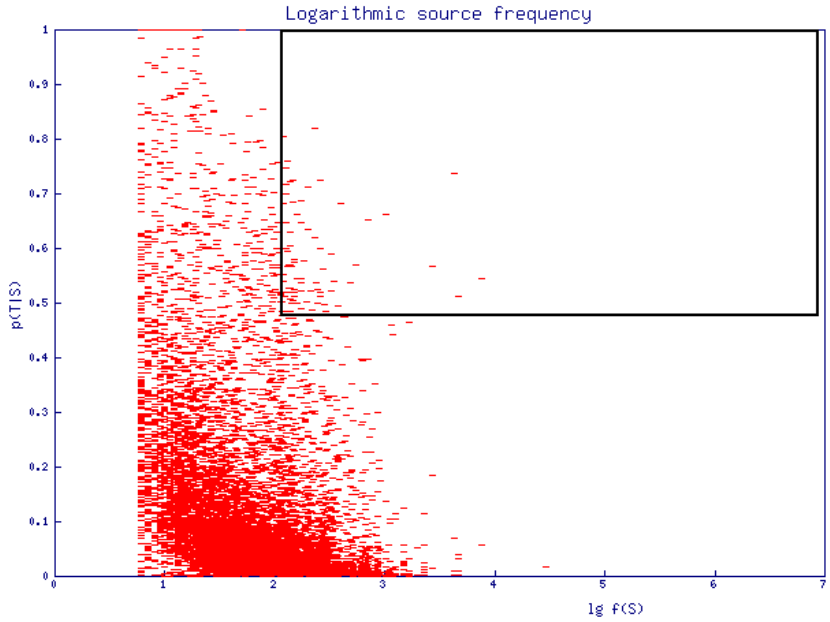
(1) A forrásnyelvi és a célnyelvi szónak is legalább 5-ször elő kell fordulnia a párhuzamos korpuszban. Ez a feltétel szükséges ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez.

(2) Hasonló gyakoriságú szavak esetén magasabb fordítási valószínűségi küszöb alkalmazása esetén magasabb lesz a jó vagy hasznos fordítási jelöltek aránya.

(3) A paraméterek beállíthatóak úgy, hogy gyakoribb forrásnyelvi szavak esetén alacsonyabb fordítási valószínűségi küszöb körülbelül ugyanolyan arányban eredményezzen jó vagy hasznos fordítási jelölteket, mint a ritkább szavak esetében egy magasabb fordítási valószínűségi küszöb.

3 Kiértékelés

Első lépésben olyan paraméterbeállítást választottunk, amely mellett feltételezhetően magas a jó vagy hasznos fordításjelöltek aránya. Így megmutathatjuk, hogy van olyan paraméterbeállítás, amely magas pontosságot eredményez, amelyből kiindulva a fedés – legalábbis részben – növelhető a paraméterbeállítások finomításával. A 2. ábrán látható a francia-holland igekeret-jelöltpárok eloszlása a forrásnyelvi kifejezés logaritmikus gyakorisága és a megfelelő fordítási valószínűség szerint. A fekete téglalap területére eső fordításjelölteket értékeltük ki. A legalább 100-szor előforduló forrásnyelvi és a célnyelvi lemmák közül azokat a fordítási jelöltpárokat választottuk ki, amelyek legalább 0,44 fordítási valószínűséggel rendelkeznek. Ezek közül 100 megfelelő keretet értékeltünk ki.



2. ábra: A francia-holland igekeret-jelöltpárok eloszlása a forrásnyelvi kifejezés logaritmikus gyakorisága és a megfelelő fordítási valószínűség szerint. A kiértékelési tartomány.

A kiértékelést két szempont alapján végeztük: egyfelől figyelembe vettük, hogy az algoritmus megtalálta-e a megfelelő igét. Másfelől azt is vizsgáltuk, hogy az illesztés a teljes keretek között történt-e. Összesen 46 esetben volt megfelelő a fordítás, úgy, hogy mind a forrásnyelvi, mind a célnyelvi oldalon teljes igei bővítménykeretek szerepeltek (46%). Ebből 54 esetben a megfelelő ige állt mindkét oldalon, de hiányos volt valamelyik, esetleg mindkét ige kerete (21 esetben a forrásnyelvi ige, 9 esetben a célnyelvi ige, 24 esetben mindkét ige kerete hiányzott).

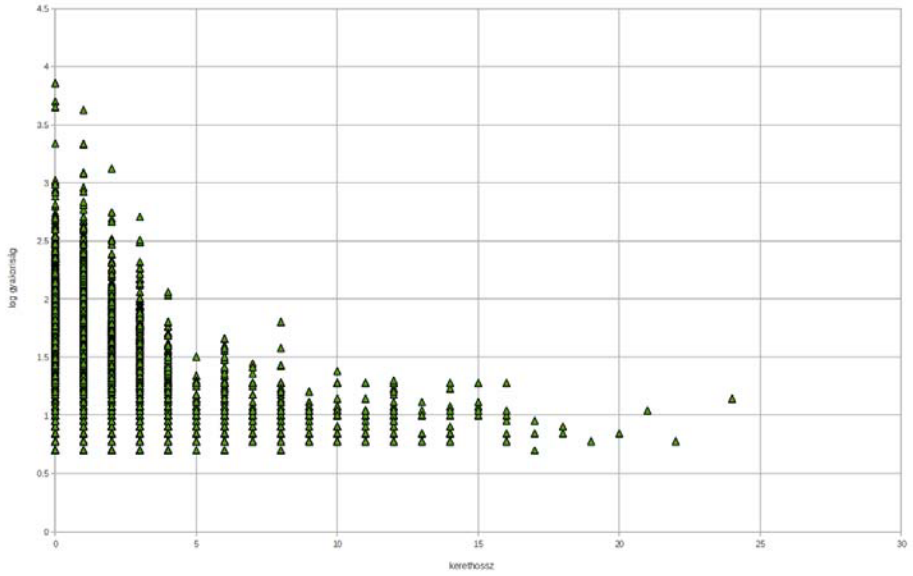
A kiértékelt keretek többnyire egy bővítményt tartalmaztak, általában egy tárgyat, de előfordultak több bővítményt tartalmazó keretek is, pl.:

avoir besoin=obj1 de:cpl *hebben obj1 nodig=predc:ADJ*
(*vkinek szüksége van vmire*)

A legjobb fordításjelöltek kiértékelése során kérdésként merült fel, hogy hogyan növelhető a jó fordításjelöltek között a teljes keretek száma? Erre egy lehetséges megoldás, hogy valamilyen alkalmas heurisztikával szűrjük a rossz kereteket az automatikusan előállított bővítménykeretlistából. Kérdés, hogy esetünkben mi számít „rossz” bővítménykeretnek. Mivel célunk általános célú szótárak építése, rossz keretnek minősülhetnek a „túl hosszú” keretek, amelyek jellemzően a korpusz valamely szaknyelvi részében (orvosi, informatikai) fordulnak elő nagy számmal. Az ilyen

keretek illesztésével a rövidebb, általánosabb kereteket kizárjuk. A leghosszabb francia keret 24 egység hosszú[2] és 14-szer fordul elő orvosi szövegekben.

A 3. ábra a francia esetében azt mutatja, hogy az egyes kerethossz alapján csoportosított kerettípusokból hány van, és az egyes keretek hányszor fordulnak elő a francia részkorpuszában.



3. ábra: A kerethossz alapján csoportosított kerettípusok száma és az egyes keretek gyakorisága a párhuzamos korpusz francia részkorpuszában.

Az ábrán jól látszik, hogy a 8 hosszúságú keretek között még vannak olyanok, amelyek viszonylag gyakoriak, így ezeket még érdemes lehet megtartani a szótár generálásánál, de az ennél hosszabbakat már nem. Mindazonáltal a keretek manuális vizsgálata azt mutatja, hogy még a 8 hosszúságú keretek is nagyon specifikusak, és egy általános célú szótár esetében legfeljebb 5 hosszúságú kereteket érdemes figyelembe venni. További empirikus vizsgálatokat igényel, hogy ez a heurisztika növeli-e a teljes keretek arányát a jó fordítási jelöltek között.

Az alkalmazott módszer érdekessége, hogy az igei szerkezetek kinyerése és a fordítási jelöltek kinyerése is felügyelet nélküli tanulással történik – vagyis az emberi intuíció kiküszöbölésével. Így a kiértékelés során azt is vizsgáltuk, hogy a kapott szerkezeteket mennyire jól karakterizálnak egy igét (*mettre*):

Az illesztés eredményeképpen előállt protoszótárból csak a 0,02-nál valószínűbb és legalább 5-ször előforduló párokat hagytuk meg. A ‘*mettre’* 5706 előfordulása 65 különböző bővítményi kerettel fordul elő. Ezek az 5611 esetben előforduló 132 holland kerettel összesen 151 fordítási párba rendeződnek. Ezeket részletesen kiértékel-

[2] A keretek hosszát a bővítmények számával mérjük: az igekinyerő algoritmusnak megfelelően a bővítmények szintaktikai funkcióját jelző morfémák és a keretben szereplő lexikai elemek ugyanolyan súllyal számítanak.

tük. A kiértékelés során igen-nem-döntést hoztunk a megfeleltetés helyességéről aszerint, hogy az adott francia keretet lehetséges-e a hozzá párosított holland kerettel fordítani a korpuszban található valamely kontextusban. Megengedtük a hiányos kereteket is, ha a konkordanciában úgy láttuk, hogy megfelelően bővíthetőek. A 151 keret 62%-át ítéltük helyesnek.

Mind a francia, mind a holland oldalon megjelöltük a hiányos kereteket, amelyek nem önálló szótári tételek, de ilyenné bővíthetőek. A '*mettre'* 65 kerete közül 10 olyan volt, amelynek csak rossz fordításai voltak, 55-höz (a keretek 85%-ához) találtunk egy vagy több helyes fordítást.

Érdekes, hogy a helytelen fordítási párok jellemzően (78% teljes francia keret és 86% teljes holland keret) a teljes keretekhez adódtak. Ezzel szemben a helyes fordítási pároknak csak 59%, illetve 63%-a teljes keret. Tehát egyértelmű trade-off van a keretek jólillesztettsége és a pontosság között.

4 Konklúziók és további teendők

Eredményeinkből látszik, hogy a javasolt módszer hasznos ötletekkel láthatja el a lexikográfusokat arra vonatkozóan, hogy mely igei tételeket kell szerepeltetni a szótárban, illetve ezen tételeknek milyen fordításai lehetnek. Mindazonáltal, a keretek sok esetben hiányosak, így sokszor kell a megfelelő konkordanciára támaszkodni a helyes igei szerkezetek visszaállításához. Így a jövőben az elsődleges célunk az, hogy a fordításjelöltek között minél teljesebb keretek szerepeljenek.

Egy lehetséges megoldás, hogy valamilyen alkalmas heurisztikával szűrjük a rossz kereteket az automatikusan előállított bővítménykeretlistából. Mivel célunk általános célú szótárak készítése, első lépésként azt kívánjuk vizsgálni, hogy a hosszú keretek rövidebb keretek alá rendezésével növelhető-e a teljes keretek aránya a fordítási jelöltpárok között.

Az eredmények általános pontosságának a növeléséhez pedig szükséges az adatok diverzitásának csökkentése, hogy minél több adat álljon az illesztő algoritmus rendelkezésére. Ehhez tovább kell szűkíteni az igeiszerkezet-algoritmus bemenetéül szolgáló nyelvtani kategóriák körét, valamint a teljes szintaktikai annotációt elegendő csak az igei szerkezeteken belül megtartani.

Bibliográfia

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
2. Bouma, G., Noord, van G., Malouf, R.: Alpino: Wide coverage computational analysis of Dutch. In: Daelemans, W., Sima'an, K., Veenstra, J., Zavrel, J. (eds): Computational Linguistics in the Netherlands 2000. Rodolpi, Amsterdam (2001) 45–59
3. 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

4. É. Kiss, K.: Mondattan. In: É. Kiss, K., Kiefer, F., Siptár, P. (eds.): Új magyar nyelvtan. Osiris Kiadó, Budapest (2003) 15–184
5. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography. In: Proceedings of the LREC2010 Conference. La Valletta, Malta (2010) 2798–2805
6. Héja E., Sass B.: Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban. In: MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2010) 80–90
7. 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)
8. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics Vol. 29, No. 1 (2003) 19–51
9. Sass, B.: A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In: Angelova G. et al. (eds.): Proceedings of RANLP 2009. Borovec, Bulgária (2009) 399–403
10. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2010) 102-110
11. Villemonte de la Clergerie: Convertir des dérivations TAG en dépendances. In: Atala, (ed.):17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010 (2010)