# Az automatikus terminológiakivonatolás módszerei és eredményei

Kis Balázs[1,2], Pohl Gábor[3]

[1] MorphoLogic Kft.
kis@morphologic.hu
[2] SZAK Kiadó Kft.
balazs.kis@szak.hu
[3] Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar
pohl@itk.ppke.hu

**Kivonat:** A terminológiakivonatolás létfontosságú mind a szakfordítási, mind pedig a terminográfiai/lexikográfiai munkában. Ennek elsősorban gazdasági jelentősége van: az általunk kifejlesztett módszerekkel jelenleg 3-6 óra alatt autentikus terminuslistához lehet jutni egy 600 oldalas (kb. 180 000 szövegszavas) szövegből, míg ugyanennek a listának a manuális előállítása 3-6 embernap munkát igényel.

Az előadás az előző évi, hasonló témájú előadás óta elvégzett kísérleteket és azok eredményeit mutatja be. Esettanulmányokon keresztül ismerteti az eddig kifejlesztett módszerek gyakorlati felhasználását.

A szakfordítás egyszerre néhány tíz–néhány száz oldalnyi (10 000–200 000 szövegszónyi) szöveggel foglalkozik. Ez erősen korlátozza a statisztikai módszerek alkalmazását, hiszen gyakorisági devianciákat, illetve asszociációs mértékeket csak jóval nagyobb korpuszokon lehet eredményesen számítani. Emiatt a terminológiakivonatolásra elsősorban szótáralapú, mintaillesztéses, illetve a legújabban környezetvizsgáló eljárásokat alkalmazunk.

## 1 A terminológiakivonatolás rendeltetése

A terminológia a szakmai nyelvhasználat elsődleges eszköze. A szakmai kommunikáció jelentős része fordításokon keresztül zajlik, ahol a fordítási folyamat sikerét elsősorban a megfelelő és konzisztens terminológiahasználat biztosítja. [12]

A fordítási folyamat résztvevői azonban nem férnek hozzá egykönnyen a terminológiához, amelynek használata a forrásnyelvi szövegben csak implicit módon jelenik meg. Emiatt, ha a nagyobb szövegek lefordítására rövid határidővel van szükség, a fordítás párhuzamosítása előtt a terminológiát elő kell készíteni [7]. A terminológia előkészítése hosszadalmas művelet, mert maximális minőségi követelmények esetén megkívánja a teljes forrásnyelvi szöveg végigolvasását. A gépi terminológiakivonatolás ezt a lépést rövidíti le jelentősen.

A gépi terminológiakivonatolás fontos szerepet tölthet be kiadványok tárgymutatóinak előkészítésében és a lexikográfiai munkában is: korpuszalapú szótárak készíté-

sekor a különböző kivonatolási eljárások a címszavak kiválasztását könnyíthetik meg. [9]

A fentiek mellett a gépi fordítás is hasznot húzhat az terminológiakivonatolásból: a fordítórendszerek által használt témaspecifikus gépi szótárak (machine-readable dictionaries; MRD) összeállításához használható, különösen akkor, ha kétnyelvű eljárások alkalmazásával a forrásnyelvi kifejezéseket és azok célnyelvi megfelelőit egyaránt előállítja. [3]

## 2 A kivonatolási eljárások

A terminológiakivonatoláshoz sokféle módszer használható, azonban nem alkalmazhatók változatlan formában azok a módszerek, amelyeket általánosan használunk többszavas kifejezések (lexémák) kivonására korpuszokból. Ennek az az oka, hogy a feldolgozandó forrásszövegek potenciálisan nem elég nagy terjedelműek ahhoz, hogy a korpusznyelvészet statisztikai eljárásai alkalmazhatóak legyenek rájuk. [6][10][11] E ponton már félreérthetetlenül látszik, hogy a fejlesztésünk tisztán gyakorlati jellegű, vagyis arra összpontosítunk, hogy eszközünk valódi fordítási feladatokban, valódi forrásnyelvi szövegeken jelentős hatékonyságnövekedést eredményezzen.

A terminológia modellezése – amelyre a terminológiakivonatolási eljárások épülnek – nem könnyű feladat. A fejlesztés során a következő alapfeltevésekből indultunk ki, amelyek a terminológiakivonatolási feladat három megközelítését nyújtják:

(1) A terminológia a szakmai nyelvhasználat alapvető attribútuma. Kis (2003) szerint a szakmai nyelvhasználat a terminológiai magatartás eredménye [7], ahol a szakmai szöveg a terminológia elemei – a terminus technicusok – mint váz köré épül. A terminológiakivonatolás feladata a váz elemeinek kielemelése.

(2) A terminusok olyan egy- vagy többszavas lexémák, amelyek a szövegben terminológiai helyzetben szerepelnek [7]. A terminológiakivonatolás feladata a szövegben szereplő szavak és kollokációk terminológiai helyzetének megállapítása vagy cáfolata.

(3) Ugyanaz a szó vagy kollokáció lehet része a szakmai szövegbeli terminológiának, más előfordulásaiazonban a diskurzust, illetve a szöveg koherenciáját biztosító nyelvi elemek közé tartozhatnak. Emellett, bár a szabványos terminológiával szemben elvárás a monoszémia, egyes terminus technicusok meghatározott szövegekben többértelműek is lehetnek (ennek oka például az lehet, hogy a szöveg több szakmai területhez is tartozik). A terminológiakivonatolás feladata, hogy megállapítsa a szöveg egyes lexémáinak (terminológiai) szerepét.

Az itt következő eljárások elsősorban a terminológiai helyzet felismerésére irányulnak, vagyis különböző kritériumokat állítanak fel arra, hogy egy adott szó vagy kollokáció terminológiai helyzetben van-e. Egyértelműsítést nem végeznek, vagyis a pontos terminológiai szerepet nem állapítják meg.

A terminológiakereső eljárások általában szabályalapúak, bár a korpusznyelvészet számos statisztikai eljárást (asszociációs mértékeket) alkalmaz kollokációk keresésére. [6][10][11] A terminológiakivonatolás bemeneti szövegei azonban általában túlságosan kis terjedelműek (< 200 000 szövegszó) ahhoz, hogy a statisztikai eljárások megbízható eredményhez vezessenek. Ugyanakkor egyes esetekben a nagy korpuszokra alkalmazott statisztikai számítások hasznosak lehetnek a szabályok előkészítésében.

### 2.1 Mintakereső eljárások

A mintakereső eljárások a terminus technicusok morfológiai-morfoszintaktikai jellemzőit próbálják megállapítani, s az ilyen jellemzőkkel rendelkező szavakat és kollokációkat keresik a szövegben. [5][7][11] A jelenlegi megvalósításunkban ezek morfoszintaktikai címkékből álló minták, amelyekben az egyes címkék felszíni sorrend szerint következnek. Példa angol nyelvű szöveg kivonatolásához használatos mintákra:

```
UNKNOWN+ADJ+N
ADJ+ADJ+N
ADJ+NUM+N
```

Ez természetesen nem a teljes mintasorozat, a jelenlegi terminológiai modellek 20-24 mintát alkalmaznak. Ez a fajta kivonatolási eljárás azonban túl sok zajt eredményez [7], ezért rendszerbe állítottunk egy heurisztikus utószűrő modult, amellyel egyes szisztematikus hibákat kerültünk el. Ezek a hibák általában gyakori, produktív kollokációkat generáló szavak formájában jelennek meg.

Az utószűrés valójában két lépésből áll:

(1) Mivel a szöveg szószintű elemzéséhez egyelőre nem egyértelműsítő szófaji jelölőt (POS-tagger), hanem morfológiai elemző programot alkalmazunk, annak eredményét is utószűrjük. Ez a szűrés feladatspecifikus, a morfológiai elemző által visszaadott egyes elemzések alkalmazását megtiltjuk.

(2) A primér mintakereső modul által visszaadott mintákat lexikális összetételük alapján szűrjük. Itt azokat a mintákat szűrjük ki, amelyek meghatározott szavakkal kezdődnek vagy végződnek.

Ezzel a módszerrel a kivonatolás pontossága (precision) kb. 20 százalékponttal javítható, ezzel 60-80%-os pontosság érhető el. Az eljárást egyelőre angol és magyar forrásnyelvű szövegek esetén alkalmaztuk. [11]

A módszert azzal fejlesztettük tovább, hogy a morfoszintaktikai minták formalizmusában eleve megengedjük lexikális korlátozások alkalmazását, hasonlóan a korábban tartalomelemzéshez alkalmazott mondatelemző rendszerhez (vö. [8]). Az utószűrést pedig teljes körűvé tettük, vagyis a javított változatban nemcsak a minták első és utolsó elemét lehet vizsgálni.

Ez utóbbi a módszerrel tovább javítható a kivonatolás pontossága, de ehhez a minták számát és bonyolultságát is növelni kell. Az elképzelt ideális felhasználói munkamódszer iteratív jellegű, vagyis a kezdeti kivonatolás után a felhasználó manuális utószűrést végez, de eközben szisztematikus szűrési utasításokat is kiad. Ezek az utasítások új utószűrési szabályok létrehozását eredményezik; emiatt olyan módszer alkalmazására lesz szükség, amellyel automatikusan is létrehozhatók ilyen szabályok.

### 2.2 Szótáras eljárások

A mintakiemelő eljárások alkalmasak ismeretlen többszavas terminus technicusok megkeresésére. Azonban célszerű kihasználni, hogy a legtöbb témakörben rendelkezünk kiinduló terminológiával, vagyis a forrásnyelvi szöveg terminológiája nem ismeretlen egészében.

Két szótáras eljárást alkalmazunk:

(1) Induktív terminológiakeresés [1][5]: ismert – szótárban tárolt – terminus technicusok összes alakjának előfordulásait keressük a forrásnyelvi szövegben, és felírjuk azon kollokációikat, amelyek megfelelnek a 2.1. fejezetben – a mintakereső eljárásoknál – leírt mintáknak. Amennyiben a mintakereső eljárások mintái kellően megengedőek, ez a szótáras eljárás mindenképpen szűkebb halmazt eredményez, mint az általános mintakeresés. A két módszer együttes alkalmazása esetén lehetőségünk van a terminusjelöltek bizonyos fokú automatikus értékelésére, mivel a szótáras induktív eljárás által megtalált terminus technicusok érvényessége valószínűbb. Magasabb pontszámot rendelhetünk azokhoz a találatokhoz, amelyek megtalálhatók voltak a szótárban, alacsonyabbat kaphatnak azok a jelöltek, amelyek egy szótári terminus technicus és egy vagy több szövegszó kollokációjából állnak, a legalacsonyabbat pedig azok a jelöltek kapják, amelyeket az általános mintakeresés eredményeként kaptunk.

(2) Az általános szókincshez nem tartozó egyszavas terminus technicusok keresése. Az eddig említett eljárások csak a többszavas terminus technicusok megkeresésére alkalmasak. Ez a szótáras eljárás azokat a szavakat keresi meg a szövegben, amelyek nem szerepelnek egy kellőképpen szűk alapszókincsben. Az alapszókincset szótárral reprezentáljuk, ez a szótár az angol és a magyar nyelvű kivonatolás esetén kb. 20 ezer címszót tartalmaz. Szakmai szövegekben valószínű, hogy az alapszókincsben nem szereplő szavak a terminológiához tartoznak, azonban sok alapszókincsbeli szó is megjelenhet terminológiai szerepben. Ez az eljárás nem alkalmas az utóbbiak felismerésére.

A második eljárás kiegészíthető kollokációkereséssel is, vagyis kereshetjük az alapszókincsben nem szereplő szavak azon kollokációit, amelyek megfelelnek a 2.1. részben – a mintakereső eljárásoknál – leírt mintáknak. Ebben az esetben az általános mintakereséssel nyert egyes jelölteket „erősíthetünk”.

A fenti eljárások közül a terminológiakivonatoló alkalmazásba egyelőre csak a másodikat integráltuk, az első eljárás egyelőre külön modulként létezik.

### 2.3 Környezetvizsgáló eljárások

A környezetvizsgáló eljárások nem a jelöltek attribútumait (belső szerkezetét), hanem a környezetük jellemzőit vizsgálják. E módszerek épülhetnek a környezet (és a jelölt) grammatikai tulajdonságaira, illetve a forrásnyelvi szöveg formázására (ha elérhető). Ilyen eljárásokat egyelőre nem implementáltunk. A lehetséges módszerek:

(1) Keressük a forrásnyelvi szövegben előforduló definíciókat, s ezek alanyát emeljük ki mint terminusjelöltet. Az eljárás megvalósításához sekély szintaktikai elemző program szükséges, amellyel egyfelől a szövegben levő főnévi csoportok határainak megkeresésére, másrészt pedig a definícióra utaló felszíni jegyek felismerésére használunk.

(2) A címek megkeresése. A szakmai szövegek belsejében szereplő címek főnévi csoportjai nagy valószínűséggel terminus technicusok, ezért egy olyan eljárás, amely felismeri a szövegbeli címeket [15], és megkeresi bennük a főnévi csoportokat, igen nagy pontosságú jelöltlistát eredményez.

### 2.4 Statisztikai módszerek: az egyszavas terminus technicusok megtalálása

Az eddig alkalmazott eljárások megfelelőnek bizonyultak többszavas terminus technicusok megkeresésére – abban az értelemben, hogy gyakorlati feladatokhoz jól használhatók, bár pontosságuk jelentősen növelhető.

Az egyszavas terminus technicusok azonban gyakran rejtve maradnak a mintakereső eljárások előtt, mivel azok szabályai közé az egyszavas mintákat általában nem vesszük fel. Így azok az egyszavas terminus technicusok, amelyek részei az alapszókincsnek, nem jelennek meg a jelöltlistán. Ezek megkeresésére alkalmazhatunk statisztikát: egyfajta „deviáns gyakoriság" módszert, amely az egyes szavak relatív gyakoriságát vizsgálja. Egyfelől szükség van egy nagy terjedelmű köznyelvi korpuszból nyert szóstatisztikára, másfelől pedig ki kell számítani a vizsgált forrásnyelvi szöveg szavainak relatív gyakoriságát. Ha egy szó relatív gyakorisága egy meghatározott küszöbértékkel meghaladja a köznyelvi korpuszbeli adatot, jó jelöltté válhat a terminus technicusok listáján.

A fenti módszer implementálása folyamatban van. Feltételezésünk szerint elsősorban 10 000 szövegszót meghaladó terjedelmű forrásnyelvi szövegeken használható majd megfelelően.

## 3 Kétnyelvű terminológiakivonatolás

A munkafolyamatot tekintve a kétnyelvű terminológiakivonatolást két lépésben valósítottuk meg:

(1) Automatikus egynyelvű terminológiakivonatolás

(2) A jelöltlista célnyelvi megfelelőinek megkeresése

A kétnyelvű terminológiakivonatolást ideális esetben szinkronizált párhuzamos szövegeken végezzük. [3][4][17] Ilyen párhuzamos szöveg a fordítómemória, amely eredeti rendeltetése szerint korábbi fordítások újrafelhasználására szolgál. Azonban a konkrét terminológiakivonatolási feladat számára – amennyiben fordítások előkészítésére használjuk – csak a forrásnyelvi szöveg ismert, mert a folyamat elvárt kimenete éppen a célnyelvi fordítás. Emiatt a kétnyelvű terminológiakivonatolás első és második lépése különböző bemeneti adatokkal működik.

A kétnyelvű terminológiakivonatolás első lépésében a forrásnyelvi szövegen egynyelvű terminológiakivonatolást végzünk, majd meghatározzuk a végleges terminuslistát. Megjegyezzük: ha a kivonatolási feladat nem fordítás előkészítésére, hanem például szótári címszavak kiválasztására, illetve szócikkek építésére szolgál, akkor az egynyelvű kivonatolás futhat párhuzamos szöveg (fordítómemória) forrásnyelvi oldalán is.

A forrásnyelvi terminuslista célnyelvi megfelelőinek meghatározása egyrészt szótárral, másrészt pedig párhuzamos szövegeken végezhető. Főnévi csoportok fordítá-

sának azonosítására alkalmazhatók már kidolgozott speciális módszerek [14]. Amennyiben rendelkezésre állnak korábbi projektekből terminológiai szótárak, egyes forrásnyelvi terminus technicusok megfelelői abban is megkereshetők. Azonban a legtöbb fordítási feladatban van új terminológia, ezért e módszer fedése sohasem 100%. A fordítási feladatokhoz ritkán készítik elő a terminológiát, ezért azokhoz fordítómemória gyakran rendelkezésre áll, korábbi terminológiai szószedet azonban alig. Ezért a piacon rendelkezésre álló fordítástámogató eszközök majdnem mindegyike nyújt konkordanciaszolgáltatást is, amely fordítómemóriák forrásnyelvi oldalán keresi meg szavak, kollokációk előfordulását, és megjeleníti az ezeket tartalmazó mondatok (szegmensek) fordítását – a konkrét terminus technicus célnyelvi pozícióját azonban már nem.

A szótáras módszer triviális, ezért a továbbiakban a célnyelvi megfelelő párhuzamos szövegből való kinyerésére összpontosítunk. A konkrét feladat olyan módszer megalkotása, amely nagy pontossággal megtalálja a forrásnyelvi terminusoknak megfelelő célnyelvi szavakat vagy kollokációkat a párhuzamos szöveg célnyelvi oldalán. Ennek előfeltétele a párhuzamos szöveg megfelelő mondatszintű szinkronizálása.

E keresés kiindulópontja valamiféle fordítási modell felállítása, amely a forrásnyelvi szavaknak célnyelvi szavakat feleltet meg. Ez általában olyan valószínűségi modell, amely annak valószínűségét határozza meg, hogy adott célnyelvi szó fordítása-e adott forrásnyelvi szónak: $P(w_T|w_S)$.

A fordítási modell lehet teljes: ez azt jelenti, hogy elvégezzük a párhuzamos szöveg teljes szószintű szinkronizálását, amelynek során a maximális $P(w_T|w_S)$ valószínűségű szópárokat keressük [3][4][14][17]. Ezeket az eljárásokat általában a statisztikai gépi fordítással összefüggésben alkalmazzák. Ebben az esetben a párhuzamos szöveg feldolgozása függetlenül végezhető az egynyelvű terminológiakivonatolástól. Az egynyelvű terminológiakivonatolás eredménylistáját a szószinten szinkronizált párhuzamos szövegen futtatjuk végig, ahol a forrásnyelvi terminus technicusok szavainak célnyelvi megfelelőit keressük ki. Ebben az eljárásban továbbra is feladat marad a többnyelvű terminus technicusok esetleg szintén többszavas fordításainak megtalálása és a megfelelő célnyelvi kifejezés (morfoszintaktikai/szintaktikai szerkezet) helyreállítása.

A rendelkezésre álló párhuzamos szöveg azonban gyakran túl kis terjedelmű ahhoz, hogy teljes fordítási modell felállításával jó minőségű eredményhez jussunk. A teljes fordítási modellre vezető algoritmusokra, illetve a statisztikai gépi fordítás eljárásaira általában is jellemző, hogy rendkívül nagy mennyiségű (több millió, több tízmillió szövegszónyi) párhuzamos szöveget igényelnek a megfelelő működéshez. Ez a mennyiség azonban a konkrét fordítási feladat esetén gyakran nem áll rendelkezésre.

A statisztikai gépi fordításban általában megengedhető, hogy különböző forrásból származó, különböző tárgykörökhöz tartozó szövegek együttes felhasználásával érjék el a „kritikus tömeget”. Ez azonban épp a terminológia meghatározása esetén ronthatja az eljárás minőségét, mert a terminológiával szemben elvárás, hogy meghatározott témakörnek, illetve szövegtípusnak megfelelően egy bizonyos fordítást alkalmazzunk, ennek megtalálására pedig nagyobb az esélyünk, ha szűrjük a kétnyelvű kivonatoláshoz felhasznált párhuzamos szövegeket.

A kétnyelvű terminológiakivonatolásban ezért „részleges” fordítási modellekkel is kísérletezünk. Ez azt jelenti, hogy a kiindulási párhuzamos szövegeink csak mondatszinten vannak szinkronizálva [13][16], és ebben keressük a forrásnyelvi (szűrt) ter-

minuslista elemeinek előfordulásait, illetve azok célnyelvi megfelelőit. A fordítási modell felállításának legfontosabb előfeltevése, hogy a terminológia fordítása konzisztens – vagyis arra számítunk, hogy a terminológiai szerepben megjelenő szavak/kifejezések fordításának lexikális összetétele mindig ugyanaz lesz. Ugyanakkor nem beszélünk a konkrét nyelvi megvalósításról, mert az mindig más lehet – ezért a forrás- és célnyelvi szegmensek szavait mindig szótő-visszaállítón és szűrőszólistán keresztül nézzük.

Árnyalja a modellt az is, hogy ugyanaz a terminus technicus – különösen, ha egyszavas – megjelenhet terminológiai helytzetben és azon kívül is, illetve interdiszciplináris szakmai szövegekben egyes terminus technicusok a terminológiai helyzetben maradva is lehetnek többértelműek. Ezért nem alkothatunk kizárólagos modellt.

Egyelőre csak kísérletek folynak e módszerek implementálására. A módszer hasonló az asszociációs mértékek számításához: azokat a célnyelvi szavakat keressük, amelyek szignifikánsan nagyobb valószínűséggel fordulnak elő olyan célnyelvi szegmensekben, amelyek forrásnyelvi oldalán a nekik megfelelő forrásnyelvi terminus technicus megtalálható. Amennyiben ez egyes szavakra nem bontható le, a célnyelvi kereshetünk két- és háromelemű kollokációkat is, amelyek esetén – mivel a terminus technicusok megfigyelésünk szerint erősen összefüggő struktúrát alkotnak – kihasználhatjuk, hogy a többszavas terminus technicusok elemei a felszínen valószínűleg szomszédosak lesznek egymással.

A fenti kísérletekre azért van szükség, mert a teljes fordítási modellek csak olyan terminus technicusok célnyelvi megfelelőinek megtalálására alkalmazhatók biztonságosan, amelyek legalább négyszer-ötször előfordulnak a forrásnyelvi szövegben. Bár a terminológiahasználat alapvető követelménye a konzisztencia, a konzisztencia pedig feltételezi az ismétlődést (tehát a többszöri előfordulást), a konkrét forrásnyelvi szövegekben a kivonatolás utáni utószűrés során elfogadott terminus technicusok 30-60%-a csak egyszer fordul elő. Mivel pedig a kétnyelvű kivonatoláshoz rendelkezésre álló párhuzamos szövegek terjedelme gyakran nem haladja meg nagyságrendelkkel a forrásnyelvi szövegét, ezért ott is nagy számban lesznek olyan terminus technicusok, amelyek a korpuszban csak egyszer fordulnak elő. Emiatt a részleges fordítási modellt érdemes szótárral támogatni, vagyis az ismert terminológiai megfeleltetéseket – a korábbi szószedeteket – felhasználni szószintű horgonyok kialakítására.

## 4 Alkalmazási példa

Az automatikus terminológiakivonatolást egy angol nyelvű szakkönyv lefordításának előkészítésére használtuk. A könyv terjedelme 151 738 szövegszó. A terminológiakivonatoláshoz olyan alkalmazást használtunk, amely a 2.1. részben, illetve a 2.2. részben leírt eljárásokat alkalmazza együtt. Az autentikus terminuslista az automatikus kivonatolás eredményének manuális utószűrésével állt elő.

A terminológiakivonatolás 12 094 jelöltet adott vissza, ebből a manuális utószűrés során 1814 (!) terminus technicust fogadtunk el. Nagyon fontos megjegyezni, hogy a manuális utószűrés eredménye nem tükrözi az eljárás pontosságát, mivel utólagos szerkesztőségi döntés alapján kb. 4000 programnyelvi kulcsszót töröltünk.

A manuális utószűrés ebben az esetben kb. 4 órát vett igénybe. Ezt az időt a könyv teljes szövegének végigolvasásához és a terminus technicusok manuális kijelöléséhez szükséges idővel kell összevetni.

A fordítási terminológia lényege azonban az, hogy a forrásnyelvi szöveg terminológiájához egyértelmű fordításokat rendel. Mivel ebben a munkában csak a forrásnyelvi szövegből nyertünk ki automatikusan terminus technicusokat, a fordítások meghatározása a manuális utómunkához tartozik. Ezt a jelen esetben a projekt terminológusa végezte, egy korábbi, hasonló témájú fordítási projekt terminológiai szótárának felhasználásával. [12]

## 5 A továbbfejlesztés irányai

Pillanatnyilag egy mintakereső és egy szótáras eljárást használunk, egy alkalmazásba integrálva. További egy szótáras eljárás megvalósítása megtörtént, a statisztikai módszer, illetve a részleges fordítási modellt alkalmazó kétnyelvű kivonatolási eljárás implementálása folyamatban van.

A továbbfejlesztés során meg kell valósítani az iteratív munkát lehetővé tevő felhasználói felületet, illetve az utószűrési szabályok (fél)automatikus generálását. Amikor pedig minden fentebb vázolt kivonatolási eljárás megvalósítása megtörtént, további kísérleteket kell végezni a pontosság növelése végett.

Az alkalmazott kivonatolási eljárások nyelvfüggetlenek, pontosabban adatvezéreltek: működésükhöz forrásnyelvi szótő-visszaállító és morfológiai elemző program (illetve, ha rendelkezésre áll, szófaji címkéző program) szükséges, emellett pedig a kivonatolási szabályokat nyelv- és néha szövegtípus-függő módon kell összeállítani. Utóbbiak azonban hozzáférhetők és szerkeszthetők a felhasználó számára.

A jövőbeli feladatok közé tartozik az is, hogy az itt kidolgozott eljárásokat további nyelvekre is kipróbáljuk.

## 6 Köszönetnyilvánítás

A szerzők szeretnének köszönetet mondani Prószéky Gábornak (MorphoLogic Kft.) a módszertani tanácsadásért, Ugray Gábornak (Kilgray Kft.) a statisztikai és az utószűrési módszerek kidolgozásában nyújtott segítségéért, Chris Callison-Burchnek (Linear B) pedig azért, hogy lehetővé tette az általa kidolgozott kereshetőfordítómemória-technológia kipróbálását.

Ez az írás közvetlen eredménye az IKTA-00181/2003. számú, a Magyar Köztársaság Oktatási Minisztériuma által támogatott projektnek.

## Bibliográfia

1. Castellví, M. T. C., Bagot, R. E. and Palatresi, J.(2001), Automatic term detection: A review of current systems, in D. Bourigault, C. Jacquemin and M.-C. L'Homme (eds), Recent Advances in Computational Terminology, John Benjamins, Amsterdam–Philadelphia, 53–88.
2. Hodász Gábor, Pohl Gábor (2005): MetaMorpho TM: a linguistically enriched translation memory. In: International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan), Borovets, Bulgaria.

3. I. Dan Melamed (2000), Models of Translational Equivalence among Words, Computational Linguistics 26(2), 221-249.
4. I. Dan Melamed (2001). Empirical Methods for Exploiting Parallel Texts. MIT Press.
5. Jacquemin, C.(2001), Spotting and Discovering Terms through Natural Language Processing, MIT Press, Cambridge (Mass.).
6. Kilgarrif, A. and Tugwell, D.(2001),Word sketch: extraction and display of significant collocations for lexicography, Proceedings of the 39th ACL and 10th EACL Workshop 'Collocation: computational extraction, analysis and explotation', Toulouse, 32–38.
7. Kis Ádám–Kis Balázs–Pohl Gábor (2004), A számítógépes terminológiakivonatolás új megközelítése. In: A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, 63-72.
8. Kis Balázs – Naszódi Mátyás – Prószéky Gábor (2003), Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. Az I. Magyar Számítógépes Nyelvészeti Konferencia előadásai, Szeged, 145-152.
9. Kis, Ádám–Kis, Balázs (2003), A prescriptive corpus-based technical dictionary. development of a multi-purpose technical dictionary, Papers in Computational Lexicography: Proceedings of COMPLEX 2003, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 47–56.
10. Kis, B., Villada, B., Bouma, G., Bíró, T., Nerbonne, J., Ugray, G. and Pohl, G. (2004), A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-Word Lexemes, Proceedings of LREC 2004, Lisbon.
11. Kis, Balázs–Villada Moirón, Begoña–Bíró, Tamás–Bouma, Gosse–Pohl, Gábor–Ugray, Gábor–Nerbonne, John (2004): Methods for the Extraction of Hungarian Multi-Word Lexemes. In: Proceedings of CLIN-2003. University of Antwerp.
12. Lengyel István–Kis Balázs–Ugray Gábor (2004), MemoQ – Új megközelítés a fordítástámogatásban. Infrastruktúratanulmány. In: A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, 100-107.
13. Pohl Gábor (2003): Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás. In: Az I. Magyar Számítógépes Nyelvészeti Konferencia előadásai, Szeged, pp 254-259.
14. Pohl Gábor (2005): Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás. In: III. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete. Szeged.
15. Pohl Gábor–Ugray Gábor (2004): Angol címek felismerése. In: II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, pp 155-160.
16. Robert C. Moore (2002): Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th AMTA Conf: Machine Translation: From Research to Real Users, pages 135-244, Langhorne, PA. Springer.
17. Chris Callison-Burch–Colin Bannard–Josh Schroeder (2005): A compact data structure for searchable translation memories. In: Practical Applications of Machine Translation. Proceedings of the 10th EAMT Conference, Pázmány Péter Catholic University, Budapest.