Kis Balázs

A fordítástechnológia és az alkalmazott nyelvtudomány

Doktori értekezés

Témavezető: dr. Prószéky Gábor

Pécsi Tudományegyetem
Nyelvtudományi Doktori Iskola
Alkalmazott Nyelvészet Program

2008. február 17.

Tartalom

Előszó

A fordítás a közelmúltban jelentős paradigmaváltáson esett át. Míg korábban leginkább egyszemélyes alkotómunkának tekintették, ma már nemigen van olyan szakmai fordítási feladat, amelyet egyetlen fordító el tudna végezni. A fordítást – mint annyi más kreatív tevékenységet – csapatok végzik, s mivel a határidők is szűkebbek lettek, számos szervezési és számítógépes eszközt kellett bevezetni ahhoz, hogy időben elvégezhetők legyenek.

A fordítás és a hozzá kapcsolódó technológia gazdasági jelentőségét mutatja, hogy a szakmai közösség számos konferenciát szervez. Ilyen a Localization World, Magyarországon az MFE által szervezett Szent Jeromos-napi találkozások vagy a legnagyobb nemzetközi fordítói szervezet, a Proz.com által szervezett összejövetelek. A fordítástechnológiáról ugyanakkor nem született tudományos igényű irodalom. Kapcsolódó területeken: a gépi fordításban, a korpusznyelvészetben, a fordítástudományban, a fordításoktatásban és a nyelvpolitikában azonban számos kutatás folyik; az irodalomjegyzékben alapvető és friss munkákat egyaránt feltüntettem. A fordítói munka műszaki, technológiai vonatkozásairól azonban nincsenek ilyen írások. A terület legfontosabbnak tekinthető forrásmunkái (Esselink 2000, Austermühl 2001) leginkább műszaki ismertetőnek, nem pedig rendszerező monográfiának tekinthetők.

Magam 1984 óta foglalkozom informatikával, 1994 óta nyelvtechnológiával. A fordítástechnológiával kimondottan gyakorlati területen találkoztam először: családi cégünk, a SZAK Kiadó olyan könyvfordítási feladatokat kapott, amelyeket az adott terjedelem és a rendelkezésre álló idő mellett nem lehetett hagyományos módszerekkel elvégezni. 1998-ban kialakítottunk egy technológiát a fordítás párhuzamosítására – különösebb gépi segédeszközök nélkül –, amely lehetővé tette a munka jó minőségű elvégzését a rendelkezésre álló idő alatt.

A speciális fordítástámogató eszközökkel a MorphoLogic munkatársaként ismerkedtem meg, ennek köszönhetően 2002 óta tanítom is ezek kezelését fordítási programok hallgatói számára. 2000-től gépi fordítással is foglalkoztam, és részt vettem a MorphoLogic MetaMorpho rendszerének kifejlesztésében is.

A SZAK Kiadó mindeközben tovább működött, így számunkra napi probléma volt a fordítások időigényének csökkentése, gazdaságosságának biztosítása – és minőségének megőrzése. Ezért érdeklődésem egyre inkább az olyan műszaki megoldások felé fordult, amelyek ebben segítettek, és doktoranduszként is elsősorban ezzel foglalkoztam. Ezek közül nem éppen a legjelentéktelenebb a fordítómemória-technológia és a terminuskivonatolás – ezekkel kapcsolatban a MorphoLogicon belül is volt alkalmam kutatás-fejlesztési projekteket vezetni.

2004-ben két munkatársammal megalapítottam a Kilgray céget, ahol kifejlesztettük a MemoQ fordítási környezetet. Ez bizonyos értelemben egyszerre

összegzése és kezdete is volt a kutatási tevékenységemnek. Érdeklődésem ugyanis eredetileg a számítógépes fordítástámogatás alkalmazott nyelvészeti vonatkozásaira irányult. Később rájöttem, hogy a fordítás gépi eszközei csak egy részét alkotják annak az eszköz- és eljárás-rendszernek, amelyet a modern fordítás igényel. A MemoQ fejlesztésének minden fázisában részt vettem, amelyek közül három terület érdekelt leginkább: a fordítók közötti hálózati együttműködés, a fordítás minőségbiztosítása és a fordítási terminológia.

Mivel a fenti kutatási tevékenység felölelte a fordítás és a fordítástechnológia minden területét, kézenfekvő volt összefoglaló jellegű doktori értekezést írni. Ennek fő célja a fordítástechnológia definiálása az alkalmazott nyelvtudomány önálló szakterületeként. A fordítástechnológia nyilvánvalóan a fordítástudománnyal, a nyelvtechnológiával és a korpusznyelvészettel áll szoros kapcsolatban, de rendkívül fontos a szociolingvisztikai, nyelvpolitikai vonatkozása is, mivel épp a fordítástechnológia jóvoltából lehet elvégezni a mai, megnövekedett terjedelmű és abszurd határidőkkel kiadott fordítási munkákat.

Az értekezés célját a fordítástechnológiai kutatások leírásával, demonstrálásával kívánja elérni, ezért rendszerszerűen, az alkalmazott nyelvtudomány különböző területeihez igazítva mutatja be kutatásokat és azok eredményeit. Az első fejezetben *definiálom a fordítástechnológiát* mint szakterületet; a második fejezetben a *nyelvpolitikai jelentőségéről* ejtek szót (Szépe 2001; Szabari 1996; Horváth 2002). Mivel úgy vélem, az alkalmazott nyelvtudomány egyetlen területétől sem választható el az oktatási tevékenység, a nyelvpolitikai fejezetben a fordítástechnológia elemeinek oktatásával is foglalkozom (Kis B. 2004, Drugan 2004). A harmadik fejezet feladata látszólag könnyű: a *fordítástudománnyal* kell kapcsolatba hoznom. A negyedik fejezetben a korpusznyelvészeté és a *nyelvtechnológiáé* a főszerep, míg az ötödik fejezet a terminológiatani vonatkozásokat tárgyalja.

A SZAK Kiadóban a fordítás technológiáját először speciális gépi eszközök nélkül alakítottuk ki – sikeresen. Ma azonban már úgy látom, hogy a fordítástechnológia működéséhez elengedhetetlen a megfelelő számítógépes eszközök felhasználása. Ezért az informatika át-átszövi az értekezést. Ugyanakkor tudatában voltam annak, hogy nem informatikai és nem is számítógépes nyelvészeti vagy korpusznyelvészeti értekezést írok, ezért az algoritmusok formális közlését és a matematikai apparátust igyekeztem a minimumra korlátozni.

Végül egy módszertani megjegyzés: az értekezésben nem definiálom az olyan alapvető fogalmakat, mint a forrásnyelv, célnyelv, forrásszöveg, célszöveg, illetve a fordítás maga. A „forrásnyelvi" és „célnyelvi" jelzőket Klaudy (2006) mintájára a legtöbbször a FNy és a CNy rövidítésekkel helyettesítem.

Köszönetnyilvánítások

A jelen értekezés alapját képező kutatásokkal 1994 óta foglalkozom (nem számítva a gimnazistaként, édesapámmal közösen fejlesztett nyelvtanulás-segítő programokat). Ez alatt az idő alatt számosan – személyek és intézmények – segítettek és motiváltak: mindannyiuknak köszönettel tartozom, különösen pedig a következőknek:

- Édesapámnak, Kis Ádámnak, akinek mind az informatika, mind a nyelvészet iránti érdeklődésemet köszönhetem, akitől rengeteget tanultam, és akivel azóta is együtt dolgozom;
- munkatársaimnak: Lengyel Istvánnak és Ugray Gábornak, akikkel közösen alapítottuk a Kilgray nevű céget fordítástechnológiai fejlesztésekre;
- a „családi cégnek": a SZAK Kiadónak, különösen édesanyámnak, Kis Ádámnénak, édesapámnak, Kis Ádámnak, és munkatársunknak, Kallósné Molnár Krisztinának, a könyvkiadó folyamatos működtetéséért: a publikációs és szótárírási lehetőségért, valamint a saját korpuszért;
- Prószéky Gábornak, akitől majdnem minden nyelvtechnológiai tudásomat és a szemléletem jó részét kaptam – és persze számos kutatási projekten dolgoztunk együtt, és könyvet is írtunk közösen;
- Szépe Györgynek, a folyamatos bátorításért és a filológia jelentőségének megismertetéséért;
- a Pécsi Tudományegyetem Alkalmazott Nyelvészeti Doktori Iskolájának, különösen Deák Péternének, azért a lehetőségért, hogy személyemben egy sok problémát okozó doktorandusz is színvonalasan végezhesse el a doktori képzést;
- az ELTE BTK Fordító- és Tolmácsképző Tanszékének, különösen Klaudy Kingának és Láng Zsuzsának, és nem utolsósorban tanítványaimnak, hogy kidolgozhattam a fordítástechnológia tanmenetét, és gyakorlati oktatásban is kipróbálhattam;
- Bach Ivánnak és Naszódi Mátyásnak, akiktől a formális nyelvek elméletét tanultam a Műegyetemen;
- a MorphoLogic munkatársainak, különösen Pál Miklósnak, Tihanyi Lászlónak, Földes Andrásnak, Endrédy Istvánnak, Novák Attilának, Aggod Andreának és Grőbler Tamásnak, akikkel számos nyelvtechnológiai projekten dolgoztunk együtt;
- a BME Automatizálási Tanszékén működő fejlesztőcsoportnak, különösen Charaf Hassannak, Juhász Sándornak és Benedek Zoltánnak, a fordítástechnológiai fejlesztésekben való együttműködésért;
- a Pázmány Péter Katolikus Egyetem Információs Technológiai Tanszékén dolgozó doktoranduszoknak: Hodász Gábornak, Miháltz Mártonnak és Pohl Gábornak, a fordítómemóriák, az információkivonatolás és a szövegszinkronizálás területén végzett közös munkáért;

- a Magyar Tudományos Akadémia Nyelvtudományi Intézete munkatársainak, különösen Váradi Tamásnak, Pajzs Júliának, Varasdi Károlynak, Gábor Katának, Oravecz Csabának, a közös kutatási projektekért, különösen a magyar mondatelemzés és a korpusznyelvészet terén;
- a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának, különösen Csirik Jánosnak, Gyimóthy Tibornak, Alexin Zoltánnak, Csendes Dórának, Hatvani Csabának a közös kutatási projektekért és a Magyar Számítógépes Nyelvészeti Konferenciáért, ahol e disszertáció témája is jelen lehet;
- a Műegyetem doktoranduszainak, Benkő Borbála Katalinnak és Katona Tamásnak a magyar mondatelemzés fejlesztésében folytatott együttes küzdelemért;
- a Miskolci Egyetem Alkalmazott Nyelvészeti Tanszékének, különösen Urbán Annának és Dobos Csillának, a fordítóképzésben való részvétel lehetőségéért;
- a Kodolányi János Főiskolának, különösen Stephanides Évának, a fordítóképzésben való részvétel lehetőségéért;
- a Szent István Egyetem Gazdasági és Társadalomtudományi Karán működő fordítóiskolának, különösen Heltai Pálnak, Dróth Júliának és Neuhauser Márknak a folyamatos szakmai együttműködésért;
- az Igazságügyi Minisztérium (volt) Fordításkoordináló Egységének, különösen Várnai Judit Szilviának és Számadó Tamásnak, a közös terminológiai munkáért;
- Voigt Vilmosnak, Pusztay Jánosnak és Kalydy Balázsnak, akik élen jártak a Magyar nyelv Terminológiai Tanácsának (MATT) megalapításában;
- a Microsoft munkatársainak, különösen Antunovics Mónikának, Barkóczi Miklósnak és Gorka Botondnak, a közös terminológiai munkáért;
- a Magyar Alkalmazott Nyelvészek Egyesületének, különösen Fóris Ágotának, a szakmai kapcsolatokért és a publikálási lehetőségekért;
- a hollandiai Rijksuniversiteit Groningen Alfa-informatica tanszéke kutatóinak, volt és jelenlegi doktoranduszainak, különösen John Nerbonne-nak, Gosse Boumának, Begoña Villada Moirónnak és Bíró Tamásnak a korpusznyelvészet terén végzett közös kutatásért;
- az European Association for Machine Translation-nek, különösen Bente Mægaardnak és John Hutchinsnak a 2005-ben rendezett budapesti EAMT-konferencia közös szervezéséért;
- az EuroTermBank-projektnek, különösen Andrejs Vasiljevsnek és Klaus-Dirk Schmitznek, a terminológiai adatbázisok fejlesztésében folytatott együttműködésért.

1. A fordítástechnológia meghatározása

1.1. A fordítástechnológia mint szakterület

A ‘technológia’ terminus alatt ebben az értekezésben nem műszaki eszközök és a hozzájuk kapcsolódó eljárások együttesét értjük. Hagyományos értelmezése szerint a technológia összetett dolgok előállításának jól definiált folyamatát írja le, magában foglalva a cél eléréséhez szükséges eszközöket, lépéseket, és ezek együttes használatának szabályait.[1]

A ‘-lógia’ szuffixum ellenére a technológia maga nem tudomány, hanem a tudomány alkalmazása. Az informatika azonban nyilvánvaló bizonyítékot szolgáltatott arra, hogy a tudomány és az alkalmazása között nem egyirányú a kapcsolat: a korpusznyelvészet például lehetővé tette a nyelv viselkedésének újszerű – kísérleti tudományhoz méltó – kutatását; a korpusznyelvészet kialakulásához pedig megfelelő teljesítményű informatikai eszközökre volt szükség.

Fordítással, nyelvészettel az ember azóta foglalkozik, amióta kommunikál. Ugyanakkor nyilvánvaló, hogy a fordítással és a nyelvvel kapcsolatos tevékenység nagy része nem tudományos, sőt, a mai szemmel nézve tudományos megközelítések csak a 19. század végén, illetve a 20. század folyamán alakultak ki.

A 20. század második felében a nyelvészetben, a nyelvhez kapcsolódó tevékenységekben, így a fordításban is megjelent a műszaki értelemben vett technika. Ez elkerülhetetlenné tette, hogy a nyelvészettel a műszaki tudományok oldaláról is foglalkozzanak: erre példa a számítógépes nyelvészet mint tudomány alkalmazása, a nyelvtechnológia. Utóbbi elsősorban a nyelvvel kapcsolatos alapvető műveletek automatizálására törekszik – e műveletek között éppen nem utolsó helyet foglal el a fordítás. Ezzel kapcsolatban két – ma már triviálisnak tűnő – megállapítást kell tennünk:

(1) A fordítást – a jelen értekezés megírásáig – nem sikerült automatizálni, amennyiben automatizálás alatt azt értjük, hogy a „fordítógép” lényeges területeken képes helyettesíteni a fordító embert. Erre bővebben kitérek a következő (1.4.) részben.

(2) A fordítás iránti kereslet napjainkra – nem: már a nyolcvanas évek végére – elérte azt a tömeget, amely mellett a fordítási feladatokat valamiféle technika igénybevétele nélkül nem lehet elvégezni.

Már legalább 20 éve léteznek olyan technikai megoldások, amelyek középutat jelentenek a teljesen individuális, ember általi fordítás és a teljesen automatikus gépi fordítás között. Az azonban nagyrészt kívül esik a módszeres vizsgálatok látókörén, hogy az említett középút – technikai eszközök alkalmazása a fordí-

táshoz kapcsolódó egyes tevékenységekben – szükségessé teszi, hogy kihasználják a rendelkezésre álló technikát használják, méghozzá meghatározott módon. Ez azt jelenti, hogy fordításhoz több, egymással együttműködő, gyakran különböző szerepeket betöltő, technikai eszközöket meghatározott szabályok szerint alkalmazó emberek együttműködése szükséges. Például ha többen fordítanak egy szakkönyvet, és rövid a határidő, a fordítást azzal lehet felgyorsítani – a minőség megőrzése mellett –, hogy a fordítók közös, hálózatban elérhető terminológiai gyűjteményt használnak, amelynek bővítését megfelelő szakember – a terminológus – felügyeli.

A fordítás mint gazdasági tevékenység rendeltetése a célnyelvi szöveg előállítása a forrásszöveg alapján. Ezt egyre ritkábban végzik egyéni fordítók: a fordítás mindinkább csapatmunka, méghozzá technikai eszközökkel segített csapatmunka lesz – ez következik az átlagos fordítási feladat nagyságából és a rendelkezésre álló időből. Arról, hogy a célnyelvi szöveg mikor tekinthető a forrásnyelvi szöveg fordításának, van közmegegyezés. A csapatmunka körülményei és az említett közmegegyezés együttesen határozzák meg azokat a szabályokat, amelyek alapján a fordítással foglalkozó embercsoportok előállítják a célnyelvi szöveget.

A fordítás ilyenformán műszaki – gyártási – tevékenységnek tekinthető, amelynek során meghatározott eszközök segítségével, meghatározott eljárások és szabályok követésével terméket állítanak elő. Ez pedig nem más, mint technológia. A fordítás olyannyira műszaki tevékenység, hogy szabványok is vonatkoznak rá, legalábbis egyes részterületeire: UNI 10574 (olasz), Önorm D 1200 és D 1201 (osztrák), DIN 2345 (német), Taalmerk (holland), ISO 12616 (nemzetközi), EN-15038 (európai). (vö. Arevalillo 2007)

A fordítástudomány vizsgálja a fordítók által követett stratégiát, vagyis törekszik annak a folyamatnak a megismerésére, amelynek során a fordítók előállítják a forrásnyelvi szöveg célnyelvi megfelelőjét. Erre elméleteket is felállít, amint arra is, hogy egy célnyelvi szöveg mikor tekinthető adott forrásnyelvi szöveg fordításának – ekvivalensének (ez utóbbiak az ekvivalenciaelméletek).

A fordítástudomány azonban mindmáig figyelmen kívül hagyja, hogy a fordítás körülményei hogyan befolyásolják a fordítás folyamatát. Ezeket a peremfeltételeket a fordítástechnológia határozza meg: a fordítás elemi művelete – egy szövegegység egy személy általi lefordítása – nagyobb rendszerbe illeszkedik, és ez a rendszer nemcsak implicit, hanem explicit módon is befolyásolja, korlátozza – vagy ha úgy tetszik, kiterjeszti – a fordító tevékenységét.

Ha a fordítás folyamatát a fordítástechnológia szemszögéből vizsgáljuk, azt mondhatjuk, hogy a fordítást végző személyek mikro- és makrostratégiákat alkalmaznak. A következőkben ezeket definiálom röviden. Ehhez vissza kell térnünk a fordítástechnológia és a számítógépes fordítástámogatás kapcsolatához.

A számítógépes fordítástámogatás szokásos megnyilvánulása a számítógépes fordítási környezetben végzett munka. Ennek során a fordító olyan

számítógépes programmal dolgozik, amely speciális – a „hagyományos" szövegszerkesztőtől eltérő – módon teszi lehetővé a fordítás megírását. Ez konkrétan a következőket jelenti:

- A fordítási környezet a forrásnyelvi szöveget automatikusan kisebb egységekre, úgynevezett szegmentumokra bontja. A legtöbb ilyen rendszerben a szegmentum leginkább a mondatnak felel meg (vagyis a gépi szegmentálás a mondathatárokat igyekszik közelíteni). A fordító egy elemi lépésben egy szegmentumot fordít le. Egyes rendszerek lehetőséget adnak a fordítónak a szegmentumok határainak módosítására – de ettől a forrásnyelvi szöveg szegmentumokra tagolása még megmarad.
- Adott szegmentum fordításához a fordítási környezet forrásokat – segítséget – ajánl fel. Ez a segítség lehet a szegmentumban előforduló egyes kifejezések fordítása (terminológia), illetve a teljes szegmentum közelítő fordítása, amennyiben a korábbi praxis során az adott szegmentumot már lefordították. Akkor is érkezhet közelítő fordítás, ha a korábbi praxisban csak az aktuális szegmentumhoz hasonló forrásszegmentumok fordítása történt meg.

Ha azt mondjuk, hogy a „technologizált fordítás" elemi művelete egy szegmentum lefordítása fordítási környezetben, akkor a fordítástechnológia folyamatai ezekből az elemi műveletekből alkotnak rendszert – akár több szinten is, hiszen a szegmentumokból előbb egy dokumentum épül fel, az pedig nagyobb rendszernek is része lehet. Megjegyezzük, hogy mivel a fordításra a gépi segítség megjelenése nyomán kezdtünk el műszaki rendszerként gondolni, azt is kijelenthetjük, hogy a fordítástechnológia kialakulását a számítógépes fordítástámogatás tette lehetővé. Maga a fordítástechnológia ugyanakkor a számítógépes fordítástámogatáshoz képest tágabb rendszer.

A fentiek alapján a fordítás mikrostratégiája az elemi művelethez, egy szegmentum lefordításához kapcsolódik. Azt határozza meg, hogy a fordító – az erőforrásoktól kapott segítséget is figyelembe véve – hogyan jut el a forrásnyelvi szegmentumtól a fordításhoz. Ugyanitt korlátozó peremfeltételek is megjelennek: amellett, hogy a fordító gondolkodását a rendszerből jövő „tippek" is befolyásolják, a fordítás során nem elegendő valamiféle általános ekvivalenciakövetelményhez alkalmazkodni. A fordítónak – ha csapat tagjaként dolgozik – igazodnia kell a csapaton belül meghatározott konzisztenciakövetelményekhez, illetve a fordítás felhasználója – megrendelője – által megkívánt formai és tartalmi követelményekhez is. Emellett a forrásnyelvi anyag formátuma is meghatározhatja a fordítás mikrostratégiáját: tipikus példa erre a szoftverhonosítás, ahol a célnyelvi szövegre méretbeli és szintaktikai korlátozások is vonatkozhatnak. A szintaktikai korlátozás alatt azt értem, hogy a honosítandó forrásnyelvi szövegben valamilyen behelyettesíthető szimbólum szerepel, amelyet a fordításban is meg kell tartani. Kézenfekvő volna a következő: *'Service %%s stopped unexpectedly.'* → **'A %%s szolgáltatás váratlanul leállt.'*, de mi történik, ha a *'%%s'* helyére magánhangzóval kezdődő szöveg kerül?

A fordítás makrostratégiája azt a folyamatot határozza meg, amelynek során megtörténik a munka előkészítése; az anyag dokumentumokra, a dokumentumok elemi szegmentumokra bontása. Ezután az elemi szegmentumok fordításából összeáll az egyes dokumentumok fordítása; megtörténik a minőségellenőrzés; végül a dokumentumok fordításából is összeáll a több dokumentumból álló célnyelvi anyag. Ennek részleteire később kitérek, most elég annyi, hogy a fordítástechnológia eszközeit és folyamatait napjainkban már valamenynyi fordítással foglalkozó szervezet kénytelen tudatosan és rendszerszerűen alkalmazni; emellett egyre kevésbé találunk olyan fordítót, aki ne lenne rendszeresen arra kényszerítve, hogy fordítástechnológiai rendszer részeként dolgozzon.

A fordítástechnológia jelentős kölcsönhatásban van az alkalmazott nyelvtudomány különböző területeivel, amellett hogy maga is a fordítástudomány egyik kutatási területe lehet.

Kutatási területként és erőforrásként a fordítástechnológia rendszerszerű kapcsolatban áll az alkalmazott nyelvtudomány több elemével:

(1) a nyelvpolitikával (a szociolingvisztikával), mivel a fordítások (megfelelő) elvégzése és léte a nyelvi jogok kérdése, sok esetben pedig jogszabály írja elő. A fordítás iránti jelenlegi kereslet, illetve a fenti körbe eső fordítási feladatok nagy volumene miatt ez csak a fordítástechnológia eszközeivel és folyamataival lehetséges, így a fordítástechnológia nyelvtervezési (korpusz- és státusztervezési) prioritást kap;

(2) a fordítástudománnyal, több oldalról is: a fordítástechnológia befolyásolja a fordítás folyamatát, egyben pedig megkönnyíti a fordítás egyes aspektusainak kutatását, egyfelől azáltal, hogy a munka során párhuzamos korpuszok jönnek létre, másfelől – a jól definiált folyamatok révén – megfigyelhető a minőségbiztosítás folyamata, és ez által további ekvivalenciamodellek állíthatók fel;

(3) a korpusznyelvészettel és azon keresztül a számítógépes nyelvészettel: a létrejövő fordítómemóriák és terminológiai adatbázisok alapanyagul szolgálnak a nyelvi elemzéssel és a gépi fordítással kapcsolatos kutatásokhoz, a lektorálás előtti és utáni szövegek összevetése pedig a fordításjavítás automatizálásának fejlesztését segíti;

(4) a terminológiatannal[2], mert a szakfordításnak fontos eleme a helyes és konzisztens terminológiahasználat. A fordítástechnológia alkalmazása szinte kizárólag a szakfordításra irányul, így fontos eleme a terminológia előkészítése, alkalmazása és ellenőrzése; a terminusok egy adott célnyelvben nagyon gyakran valamilyen fordítás által jönnek létre. A fordítástechnológiai folyamat ezért a legtöbbször egyfajta terminológiai munkafolyamatot is magában foglal.

Mivel a fordítástechnológiát az előzőekben műszaki területként is definiáltuk, interdiszciplínaként rendszerszerű kapcsolatban áll a műszaki tudományok több elemével is:

- az informatikával, azon belül a nyelvtechnológiával, mivel amellett, hogy a párhuzamos korpuszok és a gépi fordítás integrációja népszerű kutatási terület, a fordítástechnológiai rendszerek nagy mennyiségű nyelvi adat tárolását és nagy teljesítményű feldolgozását, és a meglevő párhuzamos korpuszok hatékony kihasználását igénylik. Emiatt az informatikai fejlesztés során nemtriviális adatmodellek és keresési algoritmusok kidolgozására van szükség;
- a folyamatirányítással és a projekttervezéssel, mivel a fordítástechnológiai rendszer jól definiált munkafolyamatot igényel. Azok a fordítási feladatok, amelyeket a fordítással foglalkozó szervezetek napjainkban kapnak, általában összetett projekt létrehozását igénylik.

1.2. A fordítástechnológia és a gépi fordítás

A fordítás célja, történjen bármilyen eszközzel, az emberek alapvető – ösztönös – kommunikációs igényének kielégítése, ha már a tökéletes nyelv elveszett vagy sohasem létezett. „A nyelvek összezavarodásának témája és az a törekvés, hogy az egész emberi nem közös nyelvének feltalálásával vagy felfedezésével találjanak rá gyógyírt, áthatja minden kultúra történetét." [Borst 1957, idézi Eco 1998 (1993), 17]. „Fordítás azért létezik, mert az emberek különféle nyelveket beszélnek. Legyen bár mégoly banális is ez az igazság, a helyzetről, amelyet tükröz, bízvást elmondható, hogy talányos, valamint az is, hogy súlyos pszichológiai és társadalomtörténeti kérdéseket vet fel." [Steiner 2005 (1978), 45]. Ezt számtalan módon ki lehet fejteni – Steiner [2005 (1978)] alaposan körbe is járja –, tisztán gyakorlati (nyelvpolitikai) megközelítésére a 2.1. fejezet vállalkozik.

Itt azt kell tisztázni, hogy a fordítás közvetlen rendeltetése más és más lehet, csakúgy, mint a kommunikáció szintjei. Emiatt a gépi fordítás és a géppel támogatott emberi fordítás (más szemszögből: számítógépes fordítástámogatás) rendeltetése is különbözik. Azonban érdekes lehet megfigyelni, hogy a gépi fordítás rendeltetése, illetve a két paradigma közötti „munkamegosztás" nem tudatos tervezés eredménye, hanem akcidens kutatási eredmények szerves fejlődéssel kialakult következménye. Eredete az a szemlélet, amelyet a kezdeti eufória után az ALPAC-jelentés ültetett el a kutatókban és a társadalomban: „We have already noted that while we have machine-aided translation of general scientific text, we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation."[3] (Pierce, Carroll et al. 1966:32) „For years afterwards, an interest in MT was something to keep quiet about; it was almost shameful. To this day, the «failure» of MT is still repeated by many as an indisputable fact."[4] (Hutchins 1996) Ezt megerősíti Kay

(1980) is, aki már a 80-as években definitív munkát írt az ember és a gép fordításban elfoglalt helyéről.

A fenti rébuszok megfejtése a következő: „[…] it can […] be agreed that ALPAC was quite right to be sceptical about MT: the quality was undoubtedly poor, and did not appear to justify the level of financial support it had been receiving"[5] (Hutchins 1996), tehát az ALPAC-jelentés szerint a rendelkezésünkre álló számítógépes erőforrásokkal nem lehetséges publikálható minőségű kimenetet létrehozó gépi fordítás létrehozása. (Ezt saját fejlesztési tapasztalatom is számos esetben alátámasztja.) Azért jelentett ez paradigmaváltást, mert az ALPAC-jelentést megelőzően a számítástechnikát a mesterséges intelligenciához vezető szerves – és rövid – útnak tekintették, ami alapvető ontológiai és tudományfilozófiai problémákat vet fel. Az ALPAC-jelentés azonban nem foglalkozott ilyesmivel: motivációja tisztán védelmi eredetű volt. „[…] ALPAC […] can be faulted for concentrating too exclusively on the translation needs of US scientists and of US agencies and not recognizing the broader needs of commerce and industry in an already expanding global economy."[6] (Hutchins 1996).

A jelentés önbeteljesítő jóslattá vált: nem tudhatjuk, mi történt volna, ha fenntartják a fejlesztések finanszírozását, de így az ott megfogalmazott állításokat túlnyomórészt ma is érvényesnek tekinthetjük, s közhelynek számít, hogy a jó minőségű gépi fordítás előállítása további sok évtizednyi kutatómunkát igényel. A finanszírozási türelmetlenség azóta többször is újra felszínre került, például akkor, amikor az EU a 2002-ben meghirdetett 6. keretprogram keretében már nem finanszírozta tovább a gépi fordítás fejlesztését, miközben a szervezet maga a gépi fordítás legnagyobb felhasználója. „The increased use being made of on-line machine translation demonstrates that an essentially mechanical function of that kind cannot replace the thought processes of a human translator, and thus emphasises the importance of translation quality."[7] [EC 2005:11]

A munkamegosztás alapja tehát a következő különbség: míg a gépi fordítás gyors és automatikus, rossz nyelvi minőségű és gyakran csak nagyjából érthető fordítást hoz létre, addig a géppel támogatott emberi fordítás a lényeget tekintve emberi fordítás, ezért minősége potenciálisan a lehető legjobb emberi fordítását is elérheti. Létrehozása nagyságrendekkel lassabb és több munkát igényel, mint az automatikus gépi fordítás kimenetéé, azonban – tapasztalatunk szerint – lényegesen gyorsabb, mint a gépi támogatás nélküli emberi fordítás. „Professional human translators, on the other hand, can produce good translations of many kinds of text. People can handle a range of text types; computers cannot."[8] (Melby 1995) Ez az utóbbi állítás azért érdekes, mert a gépi fordítás egyik legismertebb támogatójától (és egyben egyik legelismertebb kutatójától) származik. Ugyanő azonban kifejti azt is, hogy „The fact of the matter is that machine translation is a problem that is far from solved […]" és „a key factor […] is missing in current theories of human language […]"[9].

Itt két megjegyzést kell tennünk:

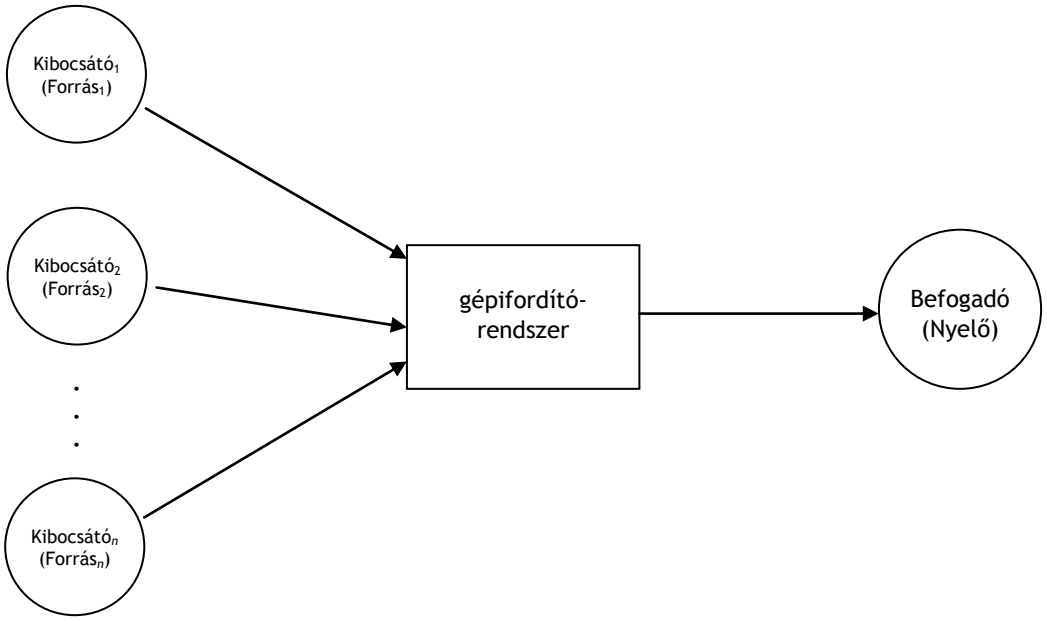
- Van két terület, ahol a gépi fordítás jó minőségű kimenetet nyújt: az egyik a kontrollált nyelvi alkalmazások területe, a másik pedig a közeli nyelvek közötti fordítás. Az előbbi olyan informatikai rendszereket jelent, amelyek korlátozzák a szöveg létrehozásához használható szókincset és grammatikai apparátust, hogy géppel hatékonyan fordítható szöveg jöjjön létre. A közeli nyelvek pedig azonos nyelvcsaládba tartozó, hasonló szókinccsel és grammatikával rendelkező nyelvek (pl. a spanyol és a katalán).

  Általában is igaz, hogy a gépifordító-rendszerek az ALPAC-jelentés megállapításainak érvényessége ellenére sokat fejlődtek az utóbbi évtizedekben, és valóban képesek jól érthető fordítások létrehozására, méghozzá sok különböző nyelvpárral.

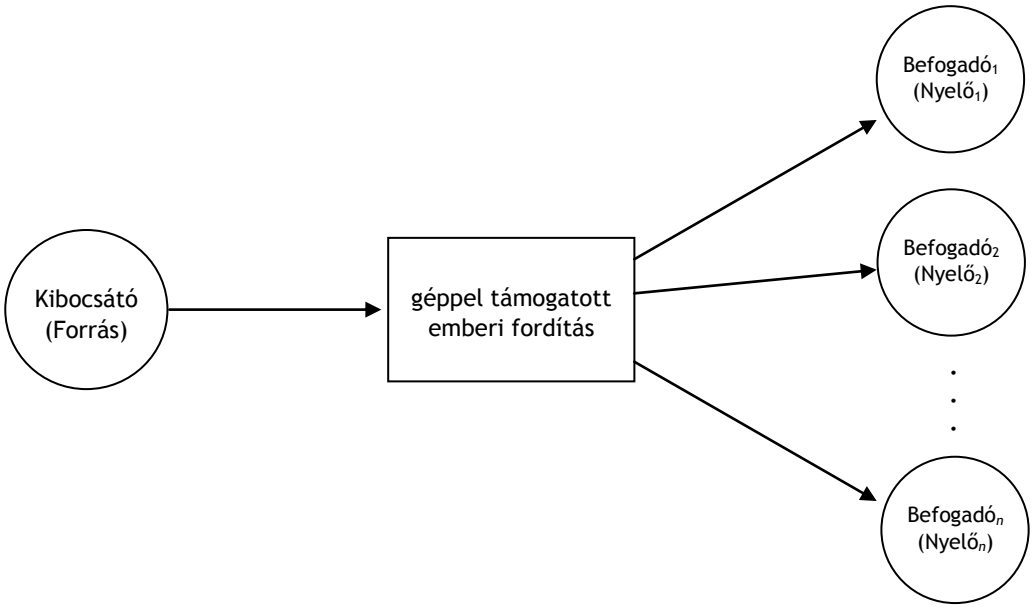
  A gépi fordítás minőségének javulásában jelentős szerepet töltött be a korpusznyelvészet megjelenése és fejlődése, amelynek pedig a számítógépek kapacitásának és sebességének növekedése volt az előfeltétele. Erről később – a 4.6. fejezetben – még lesz szó, ahol a gépi fordítás stratégiáival és minőségének mérésével is foglalkozunk.

- A fordítás minőségének értékelése nehéz feladat, s még nehezebb a minőség mérése. Alapvetően két szempontot vehetünk itt figyelembe: a nyelvi megformálás minőségét, illetve a kulturális/szemantikai ekvivalenciát. Ez a fordítástudomány területe, így a fordítástudomány feladata (lenne) a gépi vagy a gépi közreműködéssel létrehozott fordítás minőségvizsgálatára irányuló módszertan létrehozása is.

Az automatikus gépi fordítás és a géppel támogatott emberi fordítás minőségkülönbsége azt eredményezte, hogy felhasználásuk a kommunikációs lánc különböző részeihez kapcsolódik: az automatikus gépi fordítás a szöveg befogadását, míg a géppel támogatott emberi fordítás a szöveg előállítását segíti. Ez motivációbeli különbséget is jelent: a gépi fordítás felhasználását a befogadó kezdeményezi, hogy megérthessen egy számára idegen nyelven írt (vagy elmondott?) szöveget. A (géppel támogatott) emberi fordítást pedig – legalábbis áttételesen – a szöveg létrehozója használja fel, hogy mondanivalóját különböző nyelvi és helyi kultúrák számára is befogadhatóvá tegye. Ezt a következő oldal ábrái szemléltetik:



1.1. ábra. A gépi fordítás lehetséges szerepe a szöveg forrása és befogadója kapcsolatában



1.2. ábra. A (géppel támogatott) emberi fordítás lehetséges szerepe a szöveg forrása és befogadója kapcsolatában[10]

Végezetül vissza kell térnünk az ALPAC-jelentés kapcsán említett, a jelentést megelőző „utópiához". A fenti különbségtétel arra az előfeltételezésre épül, hogy a gépi intelligencia nem mérhető össze az emberi intelligenciával, illetve – egyes szemléletek szerint – nem is létezik (magam is ezt tanítom – egyelőre). Ismételve Melby (1995) állítását: „a key factor […] is missing in current theories of human language […] That key factor which is missing from current theories is agency. By agency, I mean the capacity to make real choices by exercising our will, ethical choices for which we are responsible."[11] Minden szöveg-előállítás-

sal és fordítással kapcsolatos számítógépes rendszer feltételezi, hogy az emberi intelligencia összehasonlíthatatlanul magasabb rendű, s ezért az ember által létrehozott kimenet mindig elsőbbséget élvez a gépi kimenettel szemben. A mai számítógépes rendszerek tervezési filozófiája kimondatlanul is az, hogy az emberi kimenetet mindenféle vizsgálat nélkül is jobb minőségűnek kell tekinteni a gépi kimenetnél. A jelen dolgozatban ennek különböző formáit fogjuk látni. [Holott Melby (1995) szerint: „[...] bad human translation is interesting because it was most likely done by a human yet in a manner similar to the way computers translate"[12].]

A gépi és az emberi intelligencia közötti különbség indokolására számos filozófiai (ontológiai), rendszerelméleti, matematikai és teológiai (!) érv van, ezek ismertetése azonban nem feladata ennek az írásnak: számos jelentős kutató foglalkozik ezzel. Elégedjünk itt meg a dilemma egyik megfogalmazásával [Hofstadter 1998 (1980)]:

„[…] mi, emberek, anélkül működünk, hogy ehhez *szabályokra lenne szükség:* «informális rendszerek» vagyunk. Másrészt, […] egy következtetést végző mechanikus rendszer teljes egészében szabályokra kell, hogy támaszkodjék, és emiatt nem tud elindulni, ha nincsenek olyan metaszabályai, amelyek megmondják, hogy mikor kell alkalmazni a szabályokat, meta-metaszabályok, amelyek megmondják, mikor kell alkalmazni a metaszabályokat és így tovább […].

Mi a hiba az Ördög ügyvédjének ebben a nézőpontjában? Nyilvánvalóan az a feltételezés, hogy *a gépek semmit sem képesek megtenni, ha nincs egy szabály, amely közli velük, hogy miképpen tegyék meg ezt a valamit.* [De] [...] a gépek és az emberek egyaránt olyan hardverrel rendelkeznek, amely a fizika törvényeinek engedelmeskedve teljesen magától működik. […] a legalacsonyabb szintű szabályok […] a hardverbe vannak beépítve, és anélkül futnak, hogy ehhez bármiféle engedélyre lenne szükség."

2. A fordítástechnológia nyelvpolitikai szerepe és hatása

A magyar nyelvpolitikai irodalom többnyire nem foglalkozik a fordítás nyelvtervezési vonatkozásaival, bár a szükségességét és a jelentőségét elismeri: „[…] mindez nyilván a terminológia, szaknyelv, fordítástan előtérbe kerülését eredményezheti a kutatásban és az oktatásban egyaránt" (Szépe 2001:47). Ugyanakkor általában megelégszenek az Európai Unió fordítási, tolmácsolási tevékenységének ismertetésével, anélkül hogy ebből kiindulva messzebb menő státus- és korpusztervezési vizsgálatokat végeznének (pl. Horváth 2002).

A következőkben a fordítástechnológia nyelvpolitikai vonatkozásait foglalom össze. Ez felöleli a terület kialakulásának nyelvpolitikai indítékait, a fordítótársadalomban bekövetkező változásokat, illetve a fordítástechnológia nyelvtervezési hatásait.

Ugyane fejezetben foglalkozom a gépi fordítástámogatáshoz és a fordítástechnológiával kapcsolódó oktatással is. Véleményem szerint ez szerves kapcsolatban áll a nyelvtervezéssel, mert az oktatás e tekintetben nem más, mint a társadalmi/gazdasági változások tudatos átvitele és érvényesítése.

2.1. A fordítástechnológia szükségessége

A gépi fordítástámogatás létrehozását egyaránt motiválta a névleg koordinálatlan gazdasági tevékenység és a politikai akarat. Kialakulása gazdasági szempontból kimondottan Európához, pontosabban az Európai Gazdasági Közösséghez és utódjához, az Európai Unióhoz köthető. Emögött pedig a fordításnak egy, az addigiakhoz képest új motivációját kell keresnünk, vagy legalábbis egy meglevő szempontnak a korábbiakhoz képest lényegesen nagyobb intenzitású érvényesítését.

Mindezek megértéséhez előbb át kell tekintenünk a fordítás általános motivációját. Korábban említettük, a fordítás az ember alapvető kommunikációs szükségletének kielégítésére szolgál. Elsősorban makroökonómiai szempontok miatt triviálisnak tekinthető, hogy az emberiség történelmének minden szakaszában szükségszerű volt a különböző nyelvű és kultúrájú csoportok közötti kommunikáció, így a tolmácsolás és a fordítás jóformán egyidős az emberi társadalommal, de legalábbis az írásbeliséggel. A legszebb példa erre a rosette-i kő.[13]

Az már más kérdés, hogy a különböző emberi nyelvek miért nem érthetők egymás számára, illetve az egymástól fizikailag és/vagy eredetük szerint távol első csoportok gondolkodása miért alapul gyökeresen különböző fogalmi kereteken. Erre nincs koherens válaszunk: Steiner [2005 (1978)] érdekes történeti elemzést ad róla. Ugyanakkor tudjuk, hogy az ember számára mindenféle gaz-

dasági-társadalmi érdektől függetlenül is fontos ennek a körülménynek a megszüntetése, a kölcsönös érthetőség, a „tökéletes nyelv", illetve legalábbis egy mindenki számára könnyen hozzáférhető közvetítőnyelv megtalálása [vö. Eco 1998 (1993) és Horváth 2002:3.4]. Számos mesterséges nyelv – köztük kitüntetett helyen az eszperantó – létrehozásának is ez az idealista–metafizikus megközelítés volt a motívuma.

A nemzetközi szervezetek működésének alapfeltétele a kölcsönös érthetőség. Ezt a XX. század második feléig minden nemzetközi szervezet úgy oldotta meg, hogy választott *egy* hivatalos nyelvet, amely közvetítőnyelvként működött, működik a szervezet különböző nyelvi-kulturális hátterű tagjai között. Az ókori Perzsa Birodalom fénykorában a méd, a Római Birodalomban és később a keleti egyházban a görög, a nyugati egyházban a latin volt a közvetítőnyelv. Később a Népszövetség munkanyelve a francia, az ENSZ-é az angol lett. Ez azt jelenti, hogy az adott nemzetközi szervezet tárgyalásai a munkanyelven folynak, a dokumentumok a munkanyelven keletkeznek, és – egyes propagandaanyagokat leszámítva – a kommunikáció esetleges lefordítása a tagok feladata.

A II. világháború után minden téren – politikában, nyelv- és kultúrfilozófiában, a környezetről való gondolkodásban – általánossá vált annak elfogadása, hogy nincs felsőbbrendű kultúra, nyelv és faj. Ez nemcsak azt jelenti, hogy a különböző nyelvi kultúrákat egyenlőnek tekintjük, hanem azt is, hogy mindegyiküket egyformán rendkívül fontosnak tartjuk. Ebből következik, hogy a nyelvi, kulturális, biológiai diverzitás önmagában is érték, amelyet minden lehetséges eszközzel védeni kell: „Linguistic diversity is one of the European Union's defining features. Respect for the diversity of the Union's languages is a founding principle of the European Union." (EC 2004:30)[14]

Ezzel áll párhuzamban a különböző kultúrák érzékenysége annak tekintetében, hogy a saját nyelvüket, kultúrájukat legalább annyira értékesnek tekintik, mint az összes többit, és vélt vagy valós támadás, illetve más kultúrák szupremáciaigényének megjelenése esetén izolációval vagy agresszióval védekeznek. Ennek jelenléte és intenzitása különböző, de szervesen összefügg az európai nemzetállamok XIX-XX. századbeli kialakulásával. A diverzitás előtérbe helyezése pedig *minden* ilyen érzékenység *egyidejű* tiszteletben tartását megköveteli.

A különböző nyelvek elterjedése azonban a nyelvet anyanyelvként beszélő csoport múltbeli vagy jelenlegi gazdasági/politikai erejéből következik. Így vált globális közvetítőnyelvvé az angol.

Azt mondhatjuk, hogy az Európai Unió az egyetlen nemzetközi szervezet, amely nem egyetlen hivatalos nyelvet, hanem – a jelenlegi állás szerint – huszonötöt[15] választott magának: minden tagország hivatalos nyelve az Uniónak is hivatalos nyelve. Ennek indítéka fent említett etikai megfontolás. Ez olyannyira ellentmond a gazdasági realitásnak, hogy a szervezet költségvetési okokból mégis kénytelen volt úgy dönteni, hogy a dokumentumainak az egyes hivatalos nyelvekre fordítását négy közvetítőnyelv: az angol, a francia, a német vagy a spanyol egyikének beiktatásával kell elvégezni. [Horváth 2002 (3.4)]

Az Európai Unió így egyedinek tekinthető abban, hogy a szabályai szerint minden keletkező dokumentumot le *kell* fordítani minden tagország nyelvére. Ez – tekintve a közös ügyekkel kapcsolatos jogi és más dokumentumok rendkívül nagy tömegét és keletkezésük egyre növekvő intenzitását – a fordítótársadalom számára teljesíthetetlenül nagy fordítási feladatot jelent.

Ezt ma már közhelynek tekintjük, mint ahogy azt is, hogy általában a dokumentumok tömegének és keletkezésük intenzitásának növekedése legalább közvetetten a számítástechnika számlájára írható, mivel a számítógép soha nem látott mértékben megkönnyítette a szövegek megszerkesztését és publikálását. Emiatt pedig a szövegforrások (~szerzők) populációja több nagyságrenddel növekedett.

Mivel azonban ezek a közhelyek, illetve a mögöttük levő tények vezettek a gépi fordítástámogatás kialakulásához, nem maradhatnak ki ebből a dolgozatból.

Európában tehát a közvélekedés – s az ennek megfelelően kodifikált közösségi szabályok – szerint alapvető etikai kötelesség a közösségi dokumentumok hozzáférhetővé tétele minden uniós polgár számára, a dokumentumok lefordítása alapvető kötelesség akkor is, ha lehetetlen. A fordítótársadalom kapacitása véges, ezért a fenti kötelesség teljesítésének vagy feloldásának két (három?) módja van:

- Annak elérése, hogy *minden* uniós polgár magas szinten beszéljen legalább egy közösségi munkanyelvet. Ezzel kiiktatjuk a fordítási kényszert, de áthágjuk az alapját képező etikai elvet – amely pedig meglehetősen mélyre jutott a modern gazdasági/társadalmi szemléletben, mivel még a tárgyalástechnikában is hangsúlyozzák, hogy ha egy tárgyaló fél a saját anyanyelvét beszélheti, akkor ezt az összes többi tárgyaló fél számára biztosítani kell. Ugyanakkor a diverzitás előtérbe helyezése az egyéntől is egyre inkább megköveteli, hogy az Unió minél több hivatalos nyelvét beszélje az anyanyelvén kívül: „[…] the Commission has concluded that major efforts are now required to […] make sure that everyone can speak two languages as well as their mother tongue […]". (EC 2004)[16]
- Még több fordító képzése. Erre kétségkívül szükség van, azonban vannak akadályai. Egyfelől korlátos a fordítóként szóba jöhető populáció, mivel őket csak azok közül lehet választani, akik az adott forrásnyelvet magas szinten beszélik (ha elfogadjuk azt az Unió által kodifikált alapelvet, hogy mindenki csak az anyanyelvére fordíthat). Másfelől a fordítóképzés a fordítóképző intézmények kapacitásának függvénye, és ez pedig nem növelhető minden határon túl.
- Olyan megoldás keresése, amelynek révén egységnyi szöveg lefordítása minél kevesebb munkaráfordítást jelent. Ez a fordítási munkafolyamatok optimalizálását – gépesítését és rendszerbe foglalását – jelenti, vagyis a fordítástechnológia bevezetését és alkalmazását.

Mivel úgy látszik, hogy a fordítási munka mennyisége napjainkban túlnő minden határon, a fenti három tevékenységre egyformán szükség van. Ezt az Unió is felismerte, hiszen a nemzetközi szervezetek között egyedülálló abban, hogy nagy fordítói kapacitást képviselő fordítószolgálatot működtet a saját szervezetén belül, s ennek külön főigazgatósága van *(Directorate General for Translation*, DGT). Ugyancsak intézményes keretek közé került egyes gépi fordítástámogatási eszközök felhasználása is.

A forrás (kibocsátó) oldaláról induló, szervezett fordítási tevékenységet azonban nem egyedül az EU végez. A globalizáció folyamata új piacokat nyitott a szoftvergyártók számára. Mármost a szoftver az a termék, amely elsősorban szövegesen, méghozzá nagy mennyiségű szöveg felhasználásával – kommunikál a felhasználóval. Emellett a szoftver mint termék lényegesen gyorsabban változik minden más termékfajtánál. Ez rendkívüli bonyolultságának köszönhető (egy nagyobb szoftvertermék, például egy operációs rendszer, több százmillió alapvető építőelemből áll), s ezért relatíve nagy hibaarányának, de az internet mint teljesen összekapcsolt számítógép-konglomerátum adatvédelmi problémáinak is. Meg persze annak is, hogy anyagtalan jellege miatt a megváltoztatása, továbbfejlesztése viszonylag kevés anyagi erőforrás megmozgatását igényli.

Ha a szoftvergyártó olyan terméket akar forgalmazni, amely a globális piac minden területén felhasználható, akkor úgynevezett globalizált terméket kell létrehoznia. Ez azt jelenti, hogy az adott termék minden felhasználási helyen az adott terület nyelvének és kultúrájának megfelelő módon működik. A termékből ezért minden nagyobb terület számára külön kiadást kell készíteni, különben – relevancia híján – nem adható el. A területspecifikus kiadás előállításának folyamata a lokalizáció (Esselink 2000), amelynek szükségességét a szoftvergyártók az 1990-es évek elején ismerték fel. Közülük is vezető helyen áll a Microsoft, amelynek nevét azért nem lehet itt elhallgatni, mert ez a cég építette fel a világon az első teljesen szervezett módon működő, helyi erőforrásokat bevonó lokalizációs műhelyét, amelynek mind kiterjedtsége, mind nyelvi diverzitása összemérhető az Európai Unióéval, bár a lefordítandó szövegmennyiség lényegesen kevesebb – ugyanakkor van olyan szoftvertermék, amelyet 99 nyelvre fordítanak folyamatosan.

Fontos azonban tudni, hogy a lokalizációnak van egy másfajta – decentralizált – folyamata is, amely elsősorban a nyílt forráskódú szoftvertermékek honosítására irányul. Ez olyan folyamat, amelyben helyi (sokszor ad hoc) szervezetek vesznek részt, amelyek öntevékeny módon lokalizálnak egyes nyílt forráskódú, nemzetközileg hozzáférhető szoftvertermékeket. E csoportoknak nem a lefordítandó szöveg tömege okán, hanem a rendelkezésre álló anyagi erőforrások szűkössége miatt lehet szükségük a gépi fordítástámogatásra. Azonban kérdéses ennek gazdasági realitása, mivel a jelenlegi piacon sem fordítóhoz, sem jó minőségű, csoportmunkát segítő fordítástámogató eszközhöz nem lehet ingyen hozzájutni.

Idekívánkozik egy terminológiai megjegyzés: bár a honosítást a lokalizáció szinonimájaként használják, az itt leírtak fényében mégis különbséget tennék közöttük. A lokalizáció mindig a szöveg, illetve kulturális-szöveges tartalmú termék előállítójának szemszögéből történik – ez nem más, mint a globalizálás, vagyis a globális relevanciájú termék előállítása. A honosítás viszont a befogadó kultúra szempontja: azt mondhatjuk, hogy a honosítást a célterületen működő szervezetek kezdeményezik, és nem szolgálja közvetlenül a terméket kibocsátó szervezet gazdasági érdekeit.

Összefoglalva a fentieket: megállapítottuk, hogy a különböző csoportok kommunikációs igénye globális. A globális kommunikáció lokális kommunikációt, vagyis fordítást igényel. A fordítás erőforrásai végesek, ugyanakkor egyre inkább szükség van az azonnali vagy egyidejű kommunikációra, vagyis az ennek megfelelő sebességű fordításra. Mivel megfelelő minőségű fordítást egyelőre csak az embertől várhatunk, az emberi fordítás hatékonyságát kell megnövelnünk. Ennek része a gépesítés, amelynek létjogosultságát nem lehet elvitatni, azonban jelenlegi fejlettsége mellett kételkedhetünk a valódi hasznosságában.

2.2. A technologizált fordítás társadalmi-gazdasági vonatkozásai

A fordítói munka új körülményei szükségessé teszik a csoportos fordítás alkalmazását, amelynek a gépi támogatás mellett a szervezési kérdései is fontosak, és részletes tanulmányozást igényelnek.

Idézve az előző fejezet egyik összefoglaló állítását: az emberi fordítás hatékonyságát kell megnövelni. Ennek két eszköze van: a gépesítés és a szervezés, azaz együtt a technológia. Ennek szervezési kérdéseiről, illetve annak társadalmi hatásairól kell most szót ejteni.[17]

A fordítás hagyományosan individuális tevékenység, még fordítóirodában is: egy fordító egy teljes szöveggel foglalkozik, azt az elejétől a végéig lefordítja. Ez azt jelenti, hogy egy szöveg lefordítása annyi ideig tart, amennyi idő alatt egy ember le tudja fordítani: a leggyorsabb fordítók legfeljebb kb. 20, egyenként 2000 leütést tartalmazó oldalt fordítanak egy nap alatt.

E modell követése lehetetlennek bizonyul, ha a lefordítandó szöveg terjedelme nagy, a határidő pedig szűkös. Ez általában érvényes műszaki, jogi dokumentumokra, de olyan szakkönyvekre is, amelyek tartalma gyorsan elavul.

Kézenfekvőnek látszik a munka párhuzamosítása: ha a fordítást egyetlen ember nem képes a megadott határidőre elvégezni, annyi fordítót kell alkalmaznunk, amennyire szükség van, s fel kell köztük osztanunk a fordítandó dokumentumot úgy, hogy minden fordítónak lehetőleg annyi forrásszöveget adunk, amennyit a rendelkezésre álló időben le tud fordítani.

Fontos hangsúlyozni, hogy mindezek a megállapítások és javaslatok kizárólag szakfordításokra, a szakmai kommunikációra érvényesek, a műfordításokra nem – azok esetében minimális szabályozást és technologizálást tartok szükségesnek.

A szakmai szövegek fordításának párhuzamosítása esetén nem kerülhető el a gépi fordítástámogatás alkalmazása. Ezt a konzisztencia igényével indokolhatjuk.

Az emberi fordítás minőségének legrelevánsabb jelzője az ekvivalencia. Az ezzel kapcsolatos problémákra később visszatérünk. A szakmai fordítástól (de szigorúan véve a műfordítástól is) az ekvivalencia mellett elvárjuk a konzisztenciát is. A konzisztencia elsősorban a terminológiahasználat, másodsorban pedig a nyelvezet – a regiszter, a szövegszervező elemek stb. – egységessége.

Ha egyetlen ember fordít egy hosszabb szöveget, előfordulhat, hogy a fordítás vége nem konzisztens az elejével – ez akkor következik be, ha a fordítónak nem áll a rendelkezésére olyan emlékeztető információ vagy más szabályozó eszköz, amelynek segítségével összevetheti a fordítás különböző részeit. Ezt az egyszerűség kedvéért b e l s ő i n k o n z i s z t e n c i á n a k nevezem.

Ha viszont több ember fordít egy hosszabb szöveget, a fordítás szükségképpen inkonzisztens lesz: ha nem teszünk mást, csak szétosztjuk a szöveget, minden fordító a többitől függetlenül alakít ki terminológiát és regisztert. Az előbbieknek megfelelően ez a k ü l s ő i n k o n z i s z t e n c i a .

Külső inkonzisztencia *per definitionem* csak párhuzamosított fordításban keletkezik. Ezért a csoportos fordításnak fontos eleme a fordítást megelőző előkészítés, illetve a fordítást követő egységesítés. Az előkészítés a szövegspecifikus terminológia normatív és minél teljesebb kidolgozását, illetve a regiszterre és egyéb szövegezési jellemzőkre vonatkozó útmutató kidolgozását jelenti, az egységesítés pedig a konzisztencia ellenőrzését. Amennyiben a munkát kizárólag emberek végzik, az előkészítést és az egységesítést nem végezheti egynél több ember. Mivel pedig mindkét művelet során végig kell haladni a teljes forrásszövegen, a párhuzamosított fordítás alkalmasint hosszabb időt vehet igénybe, mint individuális párja. Ha mégsem, az azt jelenti, hogy mind az előkészítés, mind az egységesítés pongyolán, a minőségi követelmények enyhítésével történik, ennek pedig az a következménye – amint látszik is néhány szövegen –, hogy átmenetileg a párhuzamos módszerrel fordított szövegek konzisztenciája és általános minősége is rosszabb, mint individuálisan előállított fordításoké. Az inkonzisztencia elkerülésének fontos eleme a fordítók képzése is. A csoportos fordításra a fordítók felkészíthetők a munka előkészítése közben is, de az ideális esetben ez még a fordítóképző intézményekben megtörténik.

Az előkészítés és az egységesítés is gyorsítható gépi segítséggel. Mivel pedig e gépi eszközök – legalábbis részben befejezett kutatások képében – már rendelkezésre állnak, megvan a lehetőség arra, hogy a csoportos (technologizált) fordítás valódi alternatívájává váljon az individuális fordításnak, mind a hatékonyság, mind a minőség tekintetében.

Most azonban arra szeretném felhívni a figyelmet, hogy a csoportos fordítás és a fordítástámogató eszközök ezt kísérő megjelenése a fordítótársadalom szerkezetét és működését is alapvetően átrendezi.

A fordítási munka egyre technológiaszerűbbé változik: a relatíve rendszertelen, illetve egyéni beosztás szerint végzett fordítás a csoportos fordításban legalább három, jól elkülöníthető szakaszra osztható, és minden szakaszban különböző, jól definiált feladatot kell elvégezni. Ezt az alábbi ábra mutatja:

2.1. ábra: A csoportos fordítás folyamatának egyszerűsített lineáris modellje

Már napjainkban is megfigyelhető, hogy a fordítók – a műfordítókat leszámítva – egyre inkább csoportmunkát végeznek, ami új eszközöket igényel, és új készségeket vár el tőlük.

A csoportos fordítás centruma olyan szervezet, amelynek az a rendeltetése, hogy elvégezze a munkaszervezést, előállítsa és rendelkezésre bocsássa a fordítási erőforrásokat (jelen esetben legalább a terminológiát). A párhuzamosított fordítási munkában a fordítók akkor tudnak hatékonyan dolgozni, ha folyamatosan hozzáférnek a közös fordítási erőforrásokhoz. A csoportos fordítás dinamikus modellje szerint a közös fordítási erőforrások a fordítási fázisban is változnak, sőt, a fordítóknak lehetőségük van visszacsatolásra. Azonban alapkövetelmény, hogy a változásoknak azonnal láthatóvá kell válniuk a többi fordító (szereplő) számára is.

Ezért a csoportmunkában dolgozó fordító – függetlenül attól, hogy irodában vagy otthon fordít – mindig hálózatban dolgozik, és folyamatosan kommunikál a többiekkel.

Ez a fajta csoportmunka a fordítótársadalom jelentős része számára egyelőre idegen, bevezetésében viszont rendkívül óvatosnak és tapintatosnak kell lenni. Az elővigyázat azért is fontos, mert a fordítóirodák, amelyek a legtöbb fordítási feladatban gazdaságossági okokból felhasználják a fordítómemóriára épülő analízist, folyamatosan és jelentősen csökkentik a fordítási díjakat, ami – bár indokolható –, jelentős feszültségeket okoz az európai fordítók között. A fordítási díjak csökkenését az az adottság indokolja, hogy a fordítási igény növekedésével nem nő párhuzamosan sem a megrendelők, sem az ügynökségek tőkeereje, s ekkor a rendeltetésük betöltésére az az egyetlen lehetőségük, hogy elvárják a fordítástámogató eszközök alkalmazását, amelyek mind az ismétlődésekkel kapcsolatos munkát, mind az azzal járó költséget megtakarítják.

2.3. A fordítástechnológia szerepe a státusztervezésben

A fordítástechnológiát az Einar Haugen (1983) által bevezetett taxonómia szerint két szempontból, a státusztervezés és a korpusztervezés szempontjából vizsgálom.

A státusztervezés felvetésére igen egyszerű válaszolni: mindaddig, amíg egy tárgykörben vagy nemzetközi szervezetben biztosítják a fordítást egy adott nyelvre vagy dialektusra, adott kultúrának megfelelően, addig fenntartható a kérdéses nyelv és nyelvközösség egyenrangú státusza az adott tárgykör vonatkozásában vagy az adott nemzetközi közösségben. Egy kommunikáló közösségben tudatos tervezés eredménye is lehet az, hogy mely nyelveket tekintünk az adott közösségen belül hivatalosnak vagy legalábbis munkanyelvnek.

A hivatalos nyelvnek jogi státusza van, a munkanyelvnek nem feltétlenül, azonban mindkét esetben olyan nyelvről van szó egy közösségen belül, amelyen a közösségen belül korlátozás nélkül lehet kommunikálni (a nyelv használatának sem technikai, sem jogi akadálya nincs). Amennyiben egy politikai szempontból erős egységet képviselő nemzetközi szervezetről van szó, egy nyelv hivatalossága vagy munkanyelv volta egyben az adott nyelvet beszélő közösség többiekkel egyenrangú vagy hozzájuk képest kiemelt státuszát is jelenti.

Egy nyelv hivatalos vagy munkanyelv volta egyben megköveteli, hogy az adott nemzetközi szervezet minden dokumentuma hozzáférhető legyen az adott nyelven, ez pedig nem más, mint a fordítási tevékenység közvetlen előírása. Ha pedig a státusztervezés során *implicit* módon fordítási tevékenységet írunk elő, akkor explicit módon gondoskodnunk kell arról, hogy az megvalósítható legyen. Így a megfelelő fordítási kapacitás létrehozása és rendelkezésre bocsátása szerves része a státusztervezési folyamatnak. Ha pedig a tudatos nyelvtervezéssel foglalkozó döntéshozó úgy látja, hogy a kellő fordítási kapacitás önmagában nem áll rendelkezésre, akkor a korábbiakban ismertetett két módon – szervezéssel és gépesítéssel – gondoskodnia kell a meglevő kapacitás hatékonyságának növeléséről.

Amit a fentiekben leírtam, az a fordításszervezés és fordítástámogatás centralizált szemléletének tekinthető. Azonban a nemzetközi szervezetek centralizált nyelvtervezési tevékenységének vannak szigorú lokális előfeltételei is.

Ahhoz, hogy egy adott tárgykör szövegei megjelenhessenek egy adott nyelven, léteznie kell az illető tárgykör terminológiájának azon a nyelven. Ezért a rendszeres fordítást terminológiateremtésnek kell megelőznie vagy kísérnie. Tudatos státustervezésnek tekinthető tehát, ha egy szakmai közösség megszervezi egy tárgykör adott nyelvű terminológiájának kialakítását, de az is, ha egy fordítócsoport szisztematikus módon előkészíti a fordítási terminológiát, s ennek során új terminusokat is alkot, majd azokat konzisztens módon használja.

A fordítás szempontjából tehát lokális vagy decentralizált státustervezésnek tekintem a terminológiaalkotást – amelyet ugyan Kis Ádám inkább korpusztervezési folyamatnak tekint (Kis Á.-Kis B. 2004), azonban a státustervezésnek a

korpusztervezés nagyon sokszor eszköze. Magyarországon ezt felismerve jött létre a Magyar Nyelv Terminológiai Tanácsa (MATT), különböző kormányzati és nem kormányzati szervezetek, illetve magánszemélyek részvételével, a Magyar UNESCO Bizottság támogatásával: „Magyarországon sokrétű terminológiával kapcsolatos tevékenység folyik, ám legtöbbször egymástól elszigetelve. A most megalakult Magyar Nyelv Terminológiai Tanács (MaTT) egyik fő küldetése egy széles nemzeti/nyelvi terminológiai kontextus kialakítása. A Tanács feladatai a magyar nyelvű terminológiával kapcsolatos alap- és alkalmazott kutatások támogatása, terminológiával kapcsolatok információk gyűjtése és terjesztése, a terminológia művelésével összefüggő kapcsolatépítés, a kormány, a gazdaság és a közigazgatás intézményeivel való kapcsolattartás.” [18]

2.4. A fordítástechnológia szerepe a korpusztervezésben

Ha egy tárgykör szövegei elsősorban fordítással kerülnek egy adott nyelvbe vagy kultúrába, akkor a fordítást végző személyek és szervezetek viselik a felelősséget az adott tárgykör adott nyelvbeli korpuszáért. Ezért a fordítástechnológia alkalmazása korpusztervezési kérdés, mivel a forrásnyelven leírt fogalmi rendszer gyakran kizárólag a fordításon keresztül kerül a célnyelvi kultúrába. Ezen keresztül a fordítástechnológiához kötődő minőségbiztosítás, illetve a terminológiatervezés kap jelentőséget.

A fordítás által érintett tárgykör nyelvhasználatára elsődleges hatással van a fordítási terminológia és a fordítások minősége is. Ezért a fordítással és az adott tárgykörrel foglalkozók előtt a következő feladatok állnak:

- Tudatos és szervezett terminológiaalkotás a fordítási munkától függetlenül
- Tudatos és szisztematikus terminológiaalkotás és terminológiahasználat a fordításban
- Konzisztens és jó nyelvi minőségű (jól olvasható) fordítások előállítása

A fentiekben a fordítástechnológia eszközei és stratégiái azért kaphatnak alapvető szerepet, mert a számítógépnek elsődleges képessége a gyors és egyértelmű információkeresés, illetve az azonosságok felismerése, illetve ellenőrzése.

Az első két feladat megoldása a terminológiai adatbázisok építése, illetve ezek hálózatban való elérhetővé tétele. Ezek azonban nem használhatók fel megfelelő ergonómiai szemlélettel kialakított fordítástámogató programok nélkül, mivel a terminológiát nem egyszerűen tárolni kell, hanem a felhasználónak – a fordítónak – fel is kell kínálni, méghozzá oly módon, hogy ez a fordítás közben ne jelentsen több munkát, mint az adatbázis használata nélkül fordítani. Így a megfelelő terminológiahasználat nem kényszeríthető ki, ha

- a terminusok nincsenek világosan megjelölve a forrásszövegben,
- a fordítónak külső erőforráshoz – akár nyomtatott szótárhoz, akár külön számítógépes programhoz – kell fordulni a célnyelvi terminológia megtalá-

lásához. Ahogyan Kis Ádám fogalmaz: „a terminológia nem a szótárban jó" (Kis Á. 2002).

A számítógép segíteni tud a konzisztencia biztosításában is. A konzisztencia biztosítása azt jelenti, hogy különböző eszközökkel elérjük, hogy ugyanaz a forrásnyelvi kifejezés vagy szegmentum mindenhol ugyanúgy legyen lefordítva. Ebből a szempontból a fordítómemória alkalmazása nemcsak a korábbi fordítások újrahasznosítását, s ezzel a fordítás hatékonyságának növelését jelenti, hanem a korábbi fordításokkal való terminológiai és regiszterbeli konzisztencia biztosítását is; emellett léteznek kereskedelmi és kísérleti eszközök a konzisztencia utólagos ellenőrzésére is.

Összefoglalva a fentieket: a státusztervezés és a korpusztervezés kettős feladata egyfelől a megfelelő fordítási kapacitás, másfelől a terminológiai bázis és a fordítások konzisztenciájának biztosítása. E feladatok – ha nagy a forrásnyelvi dokumentumok keletkezésének intenzitása, és szűkös a fordítási kapacitás – nem oldhatók meg megfelelő számítógépes segítség nélkül. A jelenleg hozzáférhető technikai eszközök erre elégtelennek bizonyulhatnak, azonban léteznek olyan, kutatási fázisban levő eszközök, amelyek további segítséget nyújthatnak.

2.5. A fordítástechnológia oktatása

A fordítástechnológiához kötődő nyelvtervezés elengedhetetlen eleme a fordítástechnológia makro- és mikrofolyamatainak (makro- és mikrostratégiáinak) oktatása, mert a fordítástechnológiát nem lehet alkalmazni a szükséges eszközök kezelésére készségszinten képes fordítók és más közreműködők nélkül. Ez egyaránt jelenti a fordításhoz felhasznált műszaki eszközök alkalmazásának (a mikrostratégiának) és a fordításszervezésnek, illetve a fordítás technikai kiszolgálásának (a makrostratégiának) az oktatását is.

2002 óta oktatom a fordítástámogató eszközök kezelését fordítóképzők hallgatói számára. Ennek során kidolgoztam egy önálló tanulásra is alkalmas tananyagot. Ezek – és az alábbi elvek – saját tapasztalataim, azonban független felmérések (Drugan 2004, Fulford–Granell-Zafra 2004) igazolják az itt leírt megállapításokat, amellett, hogy az alkalmazott módszerekre jelentős hatással volt a Microsoft által a rendszergazdák képzésére kidolgozott módszertan is.

Magyarországon az első olyan fordítóképzési program, ahol az informatikai oktatás – a fordítók számítógépes segédeszközeinek megtanításával együtt – kötelező volt a hallgatók számára, az ELTE Fordító- és Tolmácsképző Központjában indított posztgraduális EU-fordítóképzés volt. Az informatikai oktatás elsősorban a gépi fordítástámogató eszközök alapműveleteinek, vagyis a fordítás mikrostratégiájával kapcsolatos technológiának az oktatását jelentette.

Amikor fordítóknak tervezünk informatikai kurzust, körültekintően fel kell mérnünk, hogy a hallgatóknak milyen tudásra, készségekre van szükségük. A keletkező tanmenet azonban általában kompromisszumos megoldás, amely

figyelembe veszi a hallgatók előismereteit, a rendelkezésre álló időt, illetve a képzés helyén meglevő műszaki feltételeket (a hallgatók által használható hardvert és szoftvert).

Magyarországon a fordítók nagy része szabadúszóként (angolul: *freelancer)* működik, és ez valószínűleg így is marad a következő 5-6 évben. A szabadúszók több szempontból is magukra vannak hagyva. Először is, nincs mögöttük olyan állandó szervezet (fordítóiroda), amely szisztematikus követelményeket támasztana velük szemben. Ennek eredményeképp nincsenek mindig tisztában azzal, milyen informatikai készségekre van szükségük; s a megfelelő készségek megszerzésében is gyakran nehézségekbe ütköznek. Másodszor pedig nem számíthatnak műszaki segítségre, vagy ha mégis, az nem lesz formális, és nem biztos, hogy mindig a rendelkezésükre áll, amikor szükség van rá.

A fentiek miatt a fordítóknak a következő informatikai készségekkel kell rendelkezniük:

- termelési készségek: nagyon fontos, hogy fordítási teljesítményük és rendelkezésre állásuk versenyképes maradjon (Austermühl 2001);
- gépi kommunikációs készségek: ezzel kiléphetnek viszonylagos elszigeteltségükből, és tudnak (virtuális) csapatban dolgozni;
- műszaki-karbantartási készségek: fontos, hogy saját számítógépes infrastruktúrájuk működését fenn tudják tartani akkor is, amikor nem számíthatnak műszaki segítségre.

A termelési készségek a következő elemekből állnak, legalábbis ami a számítógép-használatot illeti:

- Általános szövegkezelés: alapos szövegszerkesztési ismeretekre a munkaerőpiacon szinte mindenhol szükség van, s ez fokozottan igaz a fordítók esetén. A fordítási munkafolyamatban emellett visszatérő probléma a különböző formátumú szövegek kezelése. A (szabadúszó) fordítónak nemcsak arra kell képesnek lennie, hogy kevésbé ismert formátumú dokumentumból kiemelje a szöveget, hanem sokszor meg is kell őriznie az eredeti formázást. Így egyes kiadványszerkesztő alkalmazások ismerete is szükségessé válhat.
- A speciális fordítástámogató eszközök ismerete: ez a fordítómemóriákkal, a terminológiakezelő programokkal, illetve más fordítástámogató programokkal kapcsolatos ismereteket foglalja magában. Sok fordító, illetve fordítással foglalkozó szakértő az automatikus gépi fordítást is ide sorolja. Mivel ezen eszközök egyike sem tökéletes – lényegi emberi beavatkozás nélkül egyik sem tud jó minőségű fordítást előállítani –, emellett pedig mindegyikük tartalmaz több-kevesebb nyelvtechnológiát, a jól felkészült fordítónak tisztában kell lennie e programok működésével és korlátaival. Továbbá nem elég egyetlen programmal alapos gyakorlatot szerezni, mivel a piacon alapvetően különböző alkalmazások érhetők el.

- Általános kutatási segédeszközökkel kapcsolatos készségek: a fordítóknak számos hálózati erőforrás – nemcsak elektronikus szótár – áll a rendelkezésé-re: internetes keresőrendszerek, témaspecifikus tudásbázisok, terminológiai adatbázisok stb. Ezek megfelelő kiaknázásához a fordítónak alapos internetes készségekre van szükségük, és jól kell ismerniük a keresőrendszereket is.

A legfontosabb számítógépes kommunikációs készségek a következők:

- elektronikus levelezés, csevegőprogramok stb. ismerete: ezek általános, személyes kommunikációs eszközök, amelyeket fordítási feladatok fogadására és továbbítására, illetve konzultációra lehet használni.
- hálózati csoportmunka-eszközök ismerete: számos általános csoportmunka-eszköz létezik (pl. Blackboard vagy ASAP a távoktatáshoz, konzultációhoz, Lotus Notes vagy Microsoft Exchange általánosan). Emellett hálózati munkafolyamat-kezelő rendszerek is léteznek (pl. Plunet). Jelenleg még elegendő lehet, ha a fordító mindössze tud ezekről a rendszerekről, de ahogy a folyamataik szabványosodnak, további készségek is szükségessé válhatnak.

Amint korábban is említettük, a szabadúszó fordítóknak karbantartási, hibaelhárítási tudással is kell rendelkezniük, hogy ki tudják védeni számítógépük kiesését, ami végzetes lehet, ha szoros határidővel kell dolgozniuk. Bár a személyi számítógépek látszólag nem nagyon bonyolultak, az egyes problémákat – pl. a vírustámadásokat – számos lépésben lehet csak elkerülni. A szabadúszó fordítónak ismernie kell a megfelelő eszközöket (tűzfalak, víruskereső programok stb.), és képesnek kell lennie megadott műveletek végrehajtására (pl. a biztonsági frissítések rendszeres telepítésére), hogy a számítógépének jó működését fenn tudja tartani. Ezért e fordítóknak valamelyest tisztában kell lenniük az operációs rendszer és általában a számítógép működésével, hiszen a műszaki problémákat időben meg kell találni, és el kell hárítani.

A fentiekben a fordító alapvető készségeit ismertettük, vagyis azokat, amelyekre minden fordítónak szüksége van. Azonban e készségeket több szinten is lehet birtokolni, a fordító előtt álló feladattól függően. Ezek a szintek röviden a következők:

- Alapkészségek: részletes leírásuk fentebb olvasható. A fordítóképző intézményekben elérhető képzés egyelőre csak ezeket a készségeket nyújtja.
- Weboldal-lokalizáció és filmfeliratozás: ezek olyan speciális feladatok, amelyekhez az átlagosnál jobb műszaki készségek szükségesek. A sikeres weboldal-lokalizáláshoz tisztában kell lenni a HTML- és XML-dokumentumok szerkezetével, és képesnek kell lenni a lefordítható/lefordítandó részek behatárolására. Ezen kívül szükség van webszerkesztő programok ismeretére is. A filmfeliratozáshoz pedig a speciális feliratozó programok alapos ismerete kell.
- Szoftverhonosítás: ehhez már némi programozási ismeretek is szükségesek, és alaposan tisztában kell lenni a számítógépes programok szerkezetével is.

A szoftverlokalizálással foglalkozó fordítónak a lokalizációt segítő programokat is ismernie kell. (Esselink 2000; Kis B. 2002)

- A nyelvtechnológia, illetve a fordítástechnológia kutatása: ez a tudományos ambí-ciókkal is rendelkező fordítók számára fontos. Az ő számukra már Magyarországon is rendelkezésre állnak doktori programok.

Az elmúlt évek során fokozatosan kidolgozott kurzust három szempont szerint ismertetem: (1) a kiindulási pont, vagyis a hallgatók meglevő tudása; (2) a tananyag alapprioritásai; (3) a módszertan.

A hallgatók előzetes ismereteit illetően az alábbi tényekből indulhatunk ki:

- A posztgraduális fordítóképzés hallgatóinak nagy része egyetemi (nyelvszakos) bölcsészdiplomával rendelkezik.
- Bár alapszintű számítógép-kezelést már tanítanak a bölcsészkarokon, a hallgatók általában nemigen tudnak többet alapszintű szövegszerkesztésnél – emellett a legtöbbjük tud e-mailezni, esetleg az interneten keresni.

A jelenlegi fordítóképzési programok jellemzően egy szemesztert – vagy még rövidebb időt – biztosítanak a technológiai oktatás számára; ez alatt kell átadnunk a lehető legtöbb készséget. Bár az alapszintű számítógép-kezelést sokszor elégtelenül oktatják a középiskolákban és a felsőoktatásban, ebben a kurzusban – az idő szűkössége miatt – el kell várnunk az alapismereteket. Alapszintű számítógép-kezelést tehát nem tanítunk, de a kurzus során felmerülő problémákkal foglalkozunk, és szükség esetén áttekintjük az érintett ismereteket. A kurzus legnagyobb része a speciális fordítástámogató eszközökre szorítkozik. Ennek részleteire a későbbiekben visszatérek.

Ha meg is határoztuk a szükséges készségeket, továbbra is nehéz korszerű tudást adni a hallgatóknak. Az általános termelési, kommunikációs és infrastruktúrával kapcsolatos készségek nem nagyon változnak az idővel, a speciális fordítástámogató eszközök azonban alapvető új szolgáltatásokkal egészülnek ki a következő 5 év folyamán. Korábban már említettük, hogy a fordítástámogató eszközök (főképp a fordítómemóriák és a gépi fordítás) távol vannak a tökéletességtől, és a fejlesztők jelenleg is folytatnak alapkutatási tevékenységet, amely jelentős változásokat hozhat (lásd pl. Hodász G. et al. 2004, Callison-Burch et al. 2004).

Ezek a változások várhatóan a következők lesznek:

- Intelligensebb fordítómemóriák: A példaalapú gépi fordítás *(example-based translation,* EBMT) és a nyelvérzékeny fordítómemóriák egyesítésén, illetve a „hagyományos", nyelvfüggetlen fordítómemóriák hatékonyabb kihasználásán jelenleg is több fejlesztőcsoport dolgozik, beleértve magamat is (Grőbler-Hodász-Kis 2004).
- A fordítástámogató eszközökben egyre intelligensebb terminológiakivonatoló modulok jelennek meg (Jacquemin 2001) – mivelhogy az előre meg nem adott terminológia automatikus kivonatolása egyelőre inkább csak laboratóriumi környezetben létezik.

- Egyre inkább teret nyer a csoportmunka támogatása. Ez azt jelenti, hogy új eszközök jelennek meg a fordítók, illetve a fordítási projektekben közreműködő különböző szereplők hálózati kommunikációjának biztosítására. A jelen értekezés szerzője is kidolgozott két olyan – géppel támogatott – makrostratégiai elemet (részfolyamatot), amely a technológiába további minőségbiztosítási lépéseket visz be, anélkül hogy a fordítás hatékonyságát rontaná.

A fentebb leírt körülmények mellett fontos cél az is, hogy megváltoztassuk a hallgatók attitűdjét a technika irányában (Drugan 2004). A bölcsészháttérrel érkező fordítókat sok esetben még ma is technofóbnak lehet nevezni, akik képesek ugyan megtanulni a technika használatát, de idegenkednek tőle. Ilyenformán a kurzusnak arra is fel kell készítenie a hallgatókat, hogy meg tudjanak tanulni új módszereket és új technikai eljárásokat is. Ebbe beleértjük a felszerelés önálló használatát (az infrastruktúra fenntartását, a problémák megoldását), illetve a létező technikai eszközök képességeinek és korlátainak leírását.

Ami a módszertant illeti, a kurzus majdnem teljesen mellőzi a frontális megközelítést – a „majdnem" jelentését később kifejtem.

A tanmenet a fő hangsúlyt a hallgatók önállóságára helyezi. Ez azt jelenti, hogy a tantermi foglalkozás során a hallgatók mindvégig önállóan gyakorolnak, a tanári jelenlétre ahhoz van szükség, hogy a hallgatók először is könnyen hozzáférjenek a gyakorlatok anyagához. A gyakorlatok az interneten érhetők el, és olyan módon vannak megszerkesztve, hogy önálló, tanári jelenlét nélküli tanuláshoz is használhatók legyenek.

A tanári jelenlét két ponton segíti a hallgatók munkáját:

- egyéni segítségnyújtás: a konkrét gyakorlati lépések végrehajtásának segítése egy-egy hallgató esetében;
- a hallgatók frontális informálása: egyéni hallgatói kérdések nyomán, amikor a tanár úgy ítéli, hogy a kérdés mindenki érdeklődésére számot tart, a válasz rövid előadás formájában hangzik el.

Fontos, hogy a frontális módszert ad hoc jelleggel alkalmazzuk, tehát a rövid előadások is a tanteremben konkrétan felmerülő kérdésekhez kapcsolódnak. A visszatérő (várható) problémákra adott válaszokat, kisegítő leírásokat ugyanis általában megadjuk az elektronikusan elérhető írott tananyagban.

Amennyiben az oktatáshoz egy félév (12 alkalom) vagy még rövidebb idő áll rendelkezésre, a tananyag elsősorban a mikrostratégiához kapcsolódik, a hallgatókat öt alapművelet elvégzésére készíti fel:

- A fordítás megírása és a fordítómemória használata fordítási környezetben;
- Terminológia írása és felhasználása fordítás közben;
- A fordítandó dokumentumok analízise, a fordítási költség kiszámítása, illetve árajánlat készítése;
- Szövegszinkronizálás: korábbi fordításokból származó forrásszöveg-célszöveg párok bevitele fordítómemóriába;

- Címkézett formátumú szövegek (pl. weblapok) fordítása.

Amennyiben lehetőség van második félév megtartására is, a tananyag áttér a makrostratégia elemeire:

- Fordítás-előkészítés: különböző fájlformátumok kezelése, a fordítók beosztása, terminológiai előkészítés, részletes költségvetés készítése
- Csoportos fordítás: hálózati erőforrások előkészítése és használata
- Minőségbiztosítás: előkészítés (előzetes előírások), konzisztencia-ellenőrzés, páros ellenőrzés, lektorálás

A második félév feladatai összetettek, így a fenti három feladatcsoportot egyetlen, többszereplős hallgatói projekt keretében kell elvégezni. Ehhez a hallgatók 4-5 fős hallgatói csoportokat alkotnak.

Az önállóságra szoktatás azért fontos eleme a kurzusnak, mert a rendelkezésre álló időben nincs lehetőség minden tudás átadására; sőt, a szükséges tudás behatárolása is lehetetlen, mivel az eszközök köre évről évre változik. Az önálló munka nemcsak azt segíti, hogy a hallgatók kompetenciát szerezzenek a tanteremben látott eszközök használatában, hanem azt is, hogy új eszközök és új módszerek alkalmazását is megtanulják.

Az önálló tanulást úgy is megkönnyítjük, hogy a kurzus során a mikrostratégiai alapműveleteket két különböző fordítási környezetben is elvégeztetjük. Ezt azért tesszük, mert a fordítási környezetek – az értekezés írásakor – alapvetően két különböző felhasználói paradigma szerint működnek. Ez kétféle felhasználói felület alkalmazását jelenti: a fordítási környezet beépülhet „közönséges" szövegszerkesztőbe, kiegészítve annak működését, de saját szerkesztőfelületen is lehetővé teheti a fordítás megírását. Ez utóbbi általában jelentősen egyszerűsítve van egy szövegszerkesztőhöz képest, viszont alkalmasabb a feladatra (itt általában kéthasábos táblázatban vagy osztott képernyőn dolgozunk). Az első megoldás kevesebb tanulást kíván a fordítótól, mert a megszokott szövegszerkesztőben dolgozhat; a második viszont megbízhatóbb programokat, illetve – a kezelőfelület megtanulása után – lényegesen gyorsabb munkát jelent. Az első megoldásra példa az SDL Trados Translators' Workbench vagy a WordFast, a másodikra a MemoQ, az ATRIL Déjà Vu vagy az SDLX.

Úgy látom, hogy a képzést szükséges volna kiterjeszteni magasabb szakmai szintre is, azaz „haladó" szintű számítógépes képzést is célszerű lenne nyújtani. Ebbe beleértjük a szoftver- és a weboldal-lokalizációt, figyelembe véve a fordítóhallgatók sajátos követelményeit is. Az értekezés írásakor Magyarországon már indításra készen áll néhány fordítói mesterszak, ahol az ilyen jellegű képzésre várhatóan lehetőség lesz. Ennek megfelelően a fordítástechnológia oktatói közössége is fejlődik: a tananyag létrehozása kezdetben egyéni projektum volt, ma azonban már a fordítástechnológia oktatóinak kis csoportja szerkeszti egy wikirendszerben.

3. A fordítástechnológia és a fordítástudomány

Megfigyelésem szerint a fordítástudomány (angolul 'translation studies', ritkábban 'translation theory' [Pym 1996], franciául 'traductologie' [Berman 1985]) olyan, fejlődésben levő tudományág, amelynek határai még nincsenek pontosan meghatározva. Berman (1985) definíciója kevéssé specifikus, inkább a többi tudományágtól való elkülönülést hangsúlyozza:

„The awareness of translation experiences, as distinct from all objectifying knowledge not within its framework (as dealt with by linguistics, compared literature, poetics) is what I call traductologie" [Berman 1985, vö. még: Holmes 1988(1972)]

Sokak szerint az önálló tudományágak jellemzője a saját, konzisztens módon használt terminológia. Ezt nem érzem tarthatónak: az interdiszciplínák terminológiájára, a terminológia keletkezésének problémáira még később visszatérek.

Klaudy (2006) már rendszerbe helyezi a fordítástudományt: „A fordításelmélet az alkalmazott nyelvészet egyik ága, amely a fordítás folyamatát, végeredményét és funkcióját vizsgálja a fordítási szituációban résztvevő összes nyelvi és nyelven kívüli tényező figyelembevételével." Ebben a definícióban két terminus érdemel különös figyelmet: az egyik magának a tárgykörnek a megnevezése („fordításelmélet"), amely érzésem szerint a gyakorlattal próbálja szembeállítani az elméleti kutatási területet – amelyet ugyanakkor helyesebb „tudomány"-nak nevezni, mivel olyan, egyre inkább kísérleti területről van szó, amelynek különböző kérdései nyomán versengő elméletek is születtek. A másik figyelemre méltó terminus a „nyelven kívüli tényező": itt a definíció tulajdonképpen „engedélyt ad" arra, hogy a fordítást mint kutatási területet interdiszciplínaként kapcsolatba hozzuk a műszaki tudományokkal, és nem pusztán a fordítás folyamatában megjelenő számítógépes eszközök miatt.

A fordítástechnológia mind a kutatási terület kiterjesztéseként, mind pedig kutatási erőforrásként hasznos a fordítástudomány számára. A következőkben ezt mutatom meg.

Ha a fordítástechnológiát kapcsolatba akarjuk hozni a fordítástudománnyal, meg kell fogalmaznunk a fordítástudomány azon kérdéseit, amelyekkel kapcsolatban a két terület kölcsönhatásba kerülhet. Ezeket a következőképpen előlegezhetjük meg:

- Melyek a fordítás kognitív folyamatai?
- Mit jelent a fordítási ekvivalencia, és hogyan vizsgálható?
- Melyek a fordításszövegek (grammatikai, szemantikai, szövegnyelvészeti) jellemzői? Milyen kapcsolatban állnak ezek a forrásszöveg jellemzőivel?

Tisztában kell lennünk azzal, hogy a fordítástudománynak (mint annyi más tudománynak sem a saját területén) nincsenek kész válaszai a fenti kérdésekre. Elfogadott elméletek, népszerű, használható modellek léteznek, de nem rendelkezünk természettudományos igényű bizonyítékokkal. Ez általában igaz a nyelvtudomány különböző területeire is. Ebben azonban segíthet és segít is a korpusznyelvészet és a fordítástechnológia.

Erről a következő állításokat tehetjük:

1. A fordítástechnológia a fordítástudomány vizsgálatainak tárgya. Mivel a fordítástudomány egyaránt vizsgálja a fordítás kognitív folyamatait és a célnyelvi szövegek nyelvi megformálását, szükséges vizsgálnia a hatást is, amelyet a fordítástechnológia alkalmazása gyakorol ezekre. Az értekezésben felvázolom azokat a lehetséges kutatási módszereket, amelyekkel ezek a jelenségek megvizsgálhatók. Sokaknak vannak „érzései" ezzel kapcsolatban, de megfigyeltem, hogy ezek az érzések elsősorban a gépi fordítással szembeni előítéletekből, s kevésbé valós megfigyelésekből származnak. Szükséges lenne tehát a kérdés módszeres vizsgálata; azt azonban módszertanilag igencsak megnehezíti, hogy a fordítástechnológia nélkül készített fordításokról – a fordítás mint alkotómunka individuális jellege miatt – lényegesen kevesebbet tudunk.

2. A fordítástechnológia alkalmazása során létrehozott erőforrások lehetővé teszik a fordítási folyamat, a fordítási ekvivalencia, illetve a forrásszöveg fordításnyelvre gyakorolt hatásának vizsgálatát. Erre a fordítástechnológia már pusztán azáltal alkalmassá válik, hogy a fordítás individuálisnak, tulajdonképpen intimnek tekintett lépéseit is formális keretek közé szorítja – gondoljunk arra, hogy a fordítási környezet szegmentumokra bontja a FNy szöveget, és már egyetlen szegmentum kitöltése esetén sem pusztán a fordítás begépelését teszi lehetővé. Ennek a másik oldala az, hogy a fordítástechnológia módszerei regisztrálják a fordító munkáját; ha pedig a fordítástechnológiai folyamat minőségbiztosítást is alkalmaz, akkor a technológia megőrzi a fordítás, tehát a CNy szöveg minden változatát.

Ez azt is jelenti, hogy minden fordítással foglalkozó szervezet, amely a fordítási feladatokat projektszerűen, technológiai fegyelem bevezetésével hajtja végre, kénytelen foglalkozni a fenti kérdésekkel. Más kérdés, hogy a gazdasági realitások rendszerint nem motiválják ezeket a szervezeteket arra, hogy ezeket a kutatásokat módszeresen, tudományos igénnyel elvégezzék. A technologizált fordítás során létrehozott erőforrásoktól azonban már csak egy lépés a megfelelő kutatási infrastruktúra. Az értekezés implicit módon az ilyen infrastruktúra létrehozását is bemutatja.

3. A fordítástechnológia alkalmazásával új ekvivalenciamodell jött létre. Ez a modell azt a fordítást tekinti ekvivalensnek, amelyet valamely, meghatározott technológiai fegyelmet megtartó szerkesztőség közlésre elfogad. E fegyelem megtartása esetén rendelkezünk olyan korpusszal, amely egyaránt tartalmazza a fordítás első és a szerkesztőség által elfogadott (olvasószerkesztett, korrektúrázott) változatát is, így a javításokon keresztül vizsgálhatók a fordításnyelv

jellemzői, illetve a forrásnyelvi szöveg, a fordítás és a normatív (javított, közlésre elfogadott) célnyelvi szöveg összefüggései. Az értekezés következő fejezete ennek vizsgálatával foglalkozik.

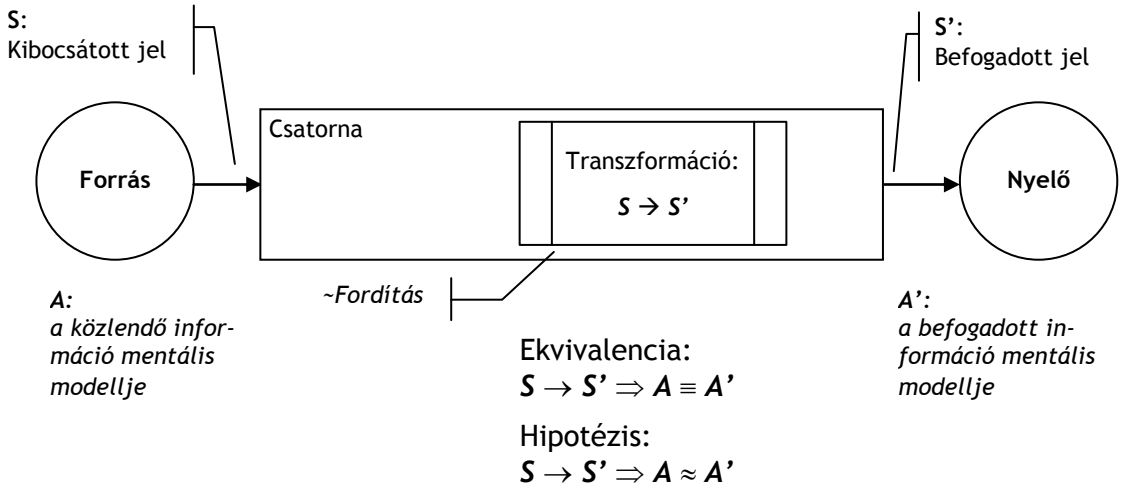
4. A fordítástechnológiai eszközök alkalmazása oly módon növeli a fordítás hatékonyságát, hogy a lényeget tekintve nem befolyásolja a fordító által alkalmazott stratégiát és módszert. Ezt a hatékonyságnövekedést kizárólag azáltal éri el, hogy elérhetővé teszi a fordításhoz szükséges információforrásokat a fordító számára. Mindez látszólagos ellentmondásban van azzal a követelménnyel (1. pont), hogy a fordítástudománynak vizsgálnia kell a fordítástechnológia hatását a fordítás folyamatára. Azonban korpuszból vett példákon keresztül az értekezés világosan bemutatja, hogy a fordítástechnológia alkalmazása sem a fordítói kreativitás (és nyelvi tudás) alkalmazásának lehetőségét, sem az iránta való igényt nem csökkenti, a konzisztencia érdekében alkalmazott formai, illetve formális korlátozások pedig lényegtelenek ebből a szempontból, pontosabban nem jelentenek nagyobb korlátozást, mint a fordítóiskolákban, fordítóirodákban és szerkesztőségekben egyébként is alkalmazott nyelvi normák.

Az értekezés bemutatja az általam társszerzőként kidolgozott fordítási környezetet, megvizsgálja az abban alkalmazható mikro- és makrostratégiát, és összeveti a fordítástechnológiához kapcsolódó, meglehetősen gyér elméleti szakirodalom szempontrendszerével.

3.1. Ekvivalencia és minőség

Az ekvivalenciaprobléma

A fordítás a kommunikáció csatornájának szerves része. Ezért az ekvivalencia problémáját is a produkció és a percepció különbségeinek figyelembe vételével kell vizsgálni.



3.1. ábra. A fordítás egyszerűsített kommunikációs modellje, rendszerelméleti jelölésekkel

A 3.1. ábra azt mutatja, hogy a kommunikációs csatorna nem feltétlenül passzív résztvevője a folyamatnak. Szigorúan véve a csatorna – legyen az átviteli közeg levegő, analóg elektronikus berendezés vagy számítógép-hálózat – mindenképp végez valamilyen jeltranszformációt. A fordítás esetében a csatornába explicit és tervezett jeltranszformátor kerül, amely nem más, mint a fordítást végző mediátor.

Ideális ekvivalenciáról akkor beszélhetünk, ha a közlés eredményeképp a befogadóban ugyanaz a mentális modell jön létre, mint amelynek közlését a kibocsátó kezdeményezte. Azonban feltételezhetjük ennek lehetetlenségét, mert az érvényes pszicholingvisztikai elméletek szerint egy közlés mentális modellje mind a kibocsátó, mind a befogadó oldalán két komponensből áll.

A befogadó oldalán ez a mentális világismeret által képviselt meglevő és a kapott jelből transzformált új információ uniója (vagy valamilyen ehhez hasonló derivátuma). A kibocsátó oldalán pedig létezik egy, a saját mentális világismeretre épülő hipotézis a befogadó mentális világismeretéről, s ehhez képest hozza létre azt a korlátozott méretű közlést – illetve először annak mentális modelljét –, amelyet végül beszélt vagy írott nyelvi jelekké alakítva továbbít a befogadó felé.

A transzformációt végző entitás – a mediátor – azonban sem a kibocsátó (forrás), sem a befogadó (nyelő) mentális világismeretét nem birtokolja, a transzformáció végrehajtásához bemenetként csak a közlést közvetítő jelet kapja meg. Ha elfogadjuk, hogy a csatornából a befogadóhoz jutó jel mindig transzformációval keletkezik, akkor biztos, hogy a csatornába bejutó és onnan kimenő jel megfelel egymásnak:

$$S' = T(S).$$

Minden jel kódolat, vagyis kódolási folyamat eredménye. Amikor a kibocsátó (forrás) a közlendő információ mentális képét jellé alakítja, kódol. A csatornában végbemenő transzformáció ezért minden esetben kódtranszformáció, amelynek felfogásom szerint három lehetséges formája (szintje) van:

(1) kódkonverzió

(2) átkódolás

(3) újrakódolás

Az első esetben a transzformáció a forrásoldali jel egyes felszíni jegyeinek megváltoztatásával állítja elő a nyelőoldali kódolatot. A második esetben (átkódolás) konverzióval egy köztes kódolatot hoz létre, s abból egy újabb konverziós lépésben a nyelőoldali kódolathoz jut. A harmadik esetben viszont teljes dekódolás történik: a mediátor a forrásoldali jel belső reprezentációját állítja elő, s a nyelőoldali kódolatot ebből a reprezentációból az előbbitől független kódolási művelettel kapja. A transzformáció műveletének paramétere a transzformációs szabályok halmaza.

Ha a fentieket az emberi fordításra vonatkoztatjuk, fel kell tételeznünk, hogy annak során a transzformáció mindig újrakódolás, amelynek paraméterei a következők:

- a mediátor (emberi fordító) saját mentális világismerete;
- a mediátor feltételezése a kibocsátó mentális világismeretéről (ennek közelítése a forrásnyelvi kultúra ismerete);
- a mediátor feltételezése a befogadó mentális világismeretéről (ennek közelítése a célnyelvi kultúra ismerete).

Ennek megfelelően a fordító dekódolás során előállítja a közlés (forrásoldali jel) mentális képét, s kódolással ebből kapja nyelőoldali jelet. Ez azonban nem mindig egyformán történik. A kutatók általában elfogadják, hogy a mentális világismeret nélkül a fordítás nem lehetséges (ennek kapcsán a gépi fordítás minőségére mint kísérleti bizonyítékra tekintenek, helytelenül[19]), viszont tudni véljük azt is, hogy a dekódolás nem feltétlenül a mentális világismeret teljes mélységében történik, s a közlés mentális képének nincs minden attribútuma kitöltve a fordítás során. Ez utóbbi hipotézist erősíti az ekvivalenciaszintek elmélete (Komisszarov 1990, idézi Klaudy 2006), amely öt kognitív szintet különböztet meg, és feltételezi, hogy a fordító az adott nyelvi elem átváltásához azt a minimális mélységű szintet választja, amelyen a művelet még éppen elvégezhető (ez tehát egy igen erős hipotézis a fordító által alkalmazott mikrostratégiára).

Ha elfogadjuk, hogy – egyelőre – sem a kibocsátó (forrás), sem a befogadó (nyelő) oldalán nincs lehetőségünk közvetlenül kinyerni a közlés A, illetve A' mentális képét, nem tudjuk ezek ekvivalenciáját egzakt módszerekkel vizsgálni. Ugyanakkor „érezzük" az ekvivalenciát, ezért kár volna kijelenteni, hogy nem létezik. De figyelem: az ekvivalencia itt semmiképpen sem egyenlőséget, inkább egyenértékűséget jelent!

Mivel a teljes (mentális, szemantikai) ekvivalencia nem vizsgálható *per se,* a kutatók az ekvivalenciát különböző szinteken próbálják megragadni (pl. Catford 1965).

A gépi fordítástámogatás szempontjából itt az a fontos, hogy amennyiben az ekvivalencia nem vizsgálható könnyen a közlés egészére nézve, meg kell keresni azokat a kisebb vagy nagyobb egységeket, amelyek esetében már formálisan is beszélhetünk ekvivalenciáról. A formálisan ábrázolt ekvivalencia azt jelenti, hogy forrásnyelvi és célnyelvi egységek között formális megfeleltetést állítunk fel, nem vizsgálva a megfeleltetés eredetét és természetét. Ugyanis ez az, amit számítógépen jól lehet ábrázolni. Ezt az ekvivalencia egyfajta közelítő modelljének tekinthetjük, de a gépi közelítések során sokszor bebizonyosodott, hogy az ilyen dekompozíció eltéríthet a lényegtől: gondoljunk például a statisztikai gépi fordításra, amelyet számosan próbálnak „enyhíteni" nem statisztikai eljárással (a találatok átrendezésével, emberi minták bevezetésével stb. – pl. Matusov et al. 2005)

Ha kézzelfogható modellt kell alkotnunk, valószínűleg akkor döntünk jól, ha egy időre félretesszük a fordítástudomány ekvivalenciaelméleteit: a literális, funkcionális és szintagmatikus ekvivalenciát, a totális fordítást (Catford 1965, Jakobson 1959), illetve a bibliafordítások által előtérbe hozott formális és dinamikus ekvivalenciát (Nida 1964). Tesszük mindezt azért, mert felismertük (a fentiekben rendszer- és kommunikációelméleti alapon is) az ekvivalenciakritériumok meghatározásának problémáit, nevezetesen (Eco 2001):

- nincs teljes szinonímia, így a lexikális egységek között sincs teljes jelentésazonosság;
- mindig van – a közlésen vagy a vizsgált közlésegységen kívüli – kontextus és konnotáció.

A fentiek mellett hiába ad Komisszarov (1990) világos besorolást az ekvivalencia különböző szintjeire, megkülönböztetve a miért, a miről, a mit és a hogyan szintjét (ezek egymásra épülnek, és rendre egyre specifikusabbak) – ezen ekvivalenciaszintek felismerése is emberi mentális kép alkotását követeli meg, részben kivéve az utolsó szintet (hogy miért, erre mindjárt visszatérünk).

Emiatt a gépi fordítástámogatás – és ma már sokszor a gépi fordítás – kutatásában másképpen tekintünk az ekvivalenciára. Szempontunkból az ekvivalencia definíciója a következő: két, különböző nyelvű közlés ekvivalens, ha ezt legalább egy, megfelelő kompetenciával rendelkező ember kijelenti. Más szavakkal: az ember által párhuzamos korpuszokba rendezett közléseket a gépi módszerek kutatói *a priori* ekvivalensnek tekintik (a későbbiekből kiderül, hogy ez legalábbis elhamarkodott feltételezés).

A párhuzamos korpuszok létrehozásával ugyanis nem oldottuk meg az alapproblémát: tudjuk, hogy a korpuszokban levő, egymásnak megfeleltetett – egymáshoz szinkronizált – közlések ekvivalensek, de nem ismerjük közelebbről az ekvivalencia természetét, pontosabban nincs szisztematikus módszerünk arra, hogy pusztán a korpuszból meg tudjuk ezt állapítani. Ismerjük a transzformáció bemenetét és kimenetét, de a transzformáció (pontosabban a mediátor) továbbra is ismeretlen.

Ebben nyújt segítséget, hogy az ekvivalencia a legkülönbözőbb strukturális szinteken vizsgálható. A számítógép tetszőleges számú és típusú szimbolikus megfeleltetést nyilván tud tartani különböző közlések között. A vizsgálatot egyelőre az írott szövegekre célszerű korlátozni, és nem foglalkozni sem a tolmácsolással, sem az interszemiotikus átalakításokkal (amelyeket Catford eleve nem is tekint fordításnak).

A párhuzamos korpuszok közötti megfeleltetések jelenleg (általában) a következők:

(1) szöveg–szöveg megfeleltetés: egymáshoz rendelünk két teljes szöveget, amelyeket egységes, de összetett közlésnek tekintünk;

(2) szakasz–szakasz megfeleltetés: a szöveg nagy strukturális egységeit feleltetjük meg egymásnak;

(3) bekezdés–bekezdés megfeleltetés: az írás szerzőjének szándéka szerinti belső tagolás egységeit feleltetjük meg egymásnak. Itt nincs garancia arra, hogy a fordítás során ez a tagolás fennmarad, de megfigyelésem szerint igen ritkán van eltérés;

(4) mondat–mondat megfeleltetés: ugyancsak az írás szerzőjének szándéka szerinti tagolás kisebb egységeit feleltetjük meg egymásnak, de a mondattagolás mint a közlés egységekre bontása sokkal inkább áll a szintaxis hatása alatt. A különböző szintaktikai formák valószínűleg különböző mértékben veszik igénybe a rövid távú memóriát, emiatt az ekvivalens közlések különböző nyelvekben időnként eltérő módon tagolódhatnak mondatokra. Erre bizonyíték lehet az a jelenség, amikor fordítás közben egy mondatot több mondatként fordítunk (az összevonás ennél ritkább);

(5) frázis–frázis megfeleltetés: a mondatnál kisebb egységek egymásnak való megfeleltetése nem mindig sikeres. Ezt a szintet ugyanis alaposan érintik a fordítási folyamatban végrehajtott átváltási műveletek, így csak annak a frázisnak a megfelelőjét találjuk meg, amelynek fordítása helyettesítéssel vagy explicitációval keletkezett;

(6) szó-szó megfeleltetés: bár egyes kutatók (pl. Callison-Burch 2005) igyekeznek ezen az alapon terminológiai szótárakat építeni szinkronizált párhuzamos korpuszokból, a szószintre leginkább a Catford-féle fordíthatatlansági hipotézis (Catford 1965) érvényes [ezt Eco (2001) is alátámasztja, amikor a teljes szinonímia lehetetlenségéről ír]. Erre teljes joggal tekinthetjük kísérleti bizonyítéknak a kizárólag szótáralapon működő gépifordító-rendszerek sikertelenségét.

A fordítási ekvivalencia vizsgálatának fontos eszközei a párhuzamos korpuszok, amelyek szisztematikus vizsgálatával statisztikai–empirikus alapon felismerhetők, illetve tömegükben vizsgálhatók a fordítók által hozott döntések, és ezeken keresztül lehet koherens ekvivalenciaelméletet kialakítani, illetve lehetővé válik a meglévő ekvivalenciaelméletek igazolása vagy cáfolása.

A párhuzamos korpusz azonban primitív, triviális ekvivalenciamodell. Azt a szemléletet tükrözi, amit a gépi fordítástámogatás, amely arra készíti fel a számítógépet, hogy kritika nélkül elfogadja és megfelelően ábrázolja az ember által kijelentett, explicite megjelölt ekvivalenciát, s ezt mintaként felhasználva megpróbálja a felszínen utánozni az emberi fordítási folyamatot. Az előbbiekben említettük, hogy az ekvivalencia mélyebb szintjeihez sem férünk hozzá. Ezért a fordítási folyamat utánzása valóban csak a felszínen, primitív formában lehetséges: a gép nem az ember átváltási műveleteit utánozza, hanem a forrásnyelvi szövegeket helyettesíti célnyelvi szövegekkel. A fordítás folyamata tehát itt is fekete doboz marad. A primitív forma a szótárak, terminológiai adatbázisok és a fordítómemóriák közvetlen felhasználását jelenti, mert ezek a fordítási folyamatot abból a kiinduló pozícióból közelítik, hogy az adott struktúrák (esetünkben konkrétan: szavak, néhány szavas kifejezések, terminusok, mondatok) meg-

feleltetései egyetemlegesek, mindig megismételhetők. Ez azonban nagyban függ az adott strukturális elem forrásszövegbeli környezetétől és a konnotációjától.

A későbbiekben bemutatok egy, a párhuzamos korpuszokra épülő, azonban már nem ennyire triviális ekvivalenciamodellt.

A fordítási ekvivalencia új modellje

A fordítástechnológia makrostratégiájának fontos eleme a minőségbiztosítás. A stratégia elemeire később még visszatérek. Azonban már most le kell szögezni, hogy a fordítási feladatok átalakulása – amely életre hívta magát a fordítástechnológiát is – két fontos következménnyel járt:

- A forrásnyelvi szöveg felosztása és a fordítás párhuzamosítása miatt szükségessé vált a fordítás konzisztenciájának biztosítása, amely teljesen egyéni munka esetén kevésbé hangsúlyozottan merül fel.
- A szűkre szabott határidő még a munka párhuzamosítása mellett sem teszi lehetővé annak a fordítói gyakorlatnak az alkalmazását, amelynek során a fordító egy ideig „pihenteti" a munkáját, majd újra áttekinti. Ezért a második személy által végzett minőség-ellenőrzésre mindenképp szükség van.

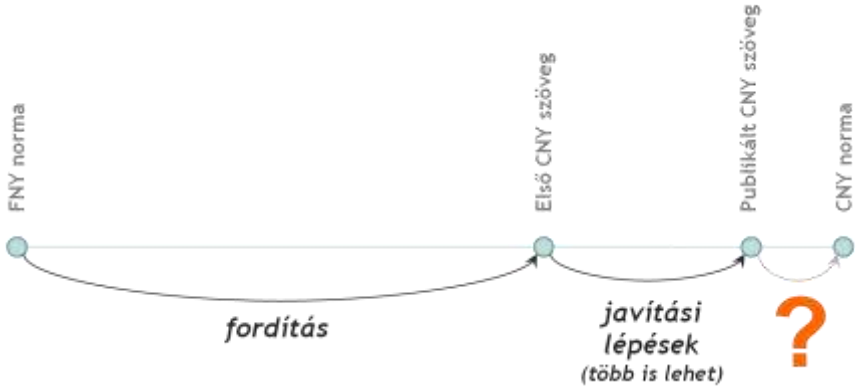
Korábban azt állítottam, hogy a számítógépes rendszerek szempontjából az ember kimenete autentikus, vagyis az ember által előállított és számítógépre mentett fordítást – az emberi kimenetet – a szoftver kritika nélkül ekvivalensnek tekinti. Azonban a fordítástechnológiai folyamatban mégiscsak jelen van a minőségbiztosítás, amely olykor két javítási fázist is jelent. A fordítással foglalkozó szervezetek (fordítóirodák, szerkesztőségek) feltételezik, hogy a javítás során a szöveg minősége „jobb" lesz, ez fordítások esetén azt is jelenti, hogy „ekvivalensebb" lesz a fordító által elsőként készített változatnál. Néhány példa:

3.2. ábra. Minták javításkorpuszból

A fordítás tényéből világosan látszik, hogy a fordítónak van mentális képe az ekvivalenciáról. A javításokból pedig az látszik, hogy a lektornak (lektoroknak) is van. Ezt a mentális képet nem ismerjük, de feltételezzük (a fordítással foglalkozó szervezetek, szerkesztőségek elfogadják), hogy a lektorálás által a szöveg közelebb kerül valamiféle ideálisan ekvivalens állapothoz. Az ekvivalenciába beleértjük azt is, hogy a CNy szöveg „ugyanúgy" illeszkedik a célnyelv rendszerébe, ahogy a FNy szöveg a forrásnyelvébe, bármit is jelentsen az „ugyanúgy" – ezért a nyelvi, helyesírási javításokat itt ugyanúgy nem szabad figyelmen kívül hagynunk, mint a tartalmiakat.

Az embernek van tehát valamilyen képe az ideálisan ekvivalens fordításról, sőt, ebben valamiféle közmegegyezés is létezik – ez abból látszik, hogy sokan igyekeznek megfogalmazni a fordítással kapcsolatos normákat, és olyan módszereket kidolgozni a fordítás oktatásához, amelyekkel e normák átadhatók, begyakorolhatók és számon kérhetők.

Egyes elméletek szerint az ekvivalencia több szinten jöhet létre. Az elmélet ezt a nyelvi struktúrához köti. Mivel azonban erre nincs kísérleti bizonyíték, óvakodnunk kell attól, hogy a modellalkotás során a megfigyeléseinket bármelyik elmélethez is hozzáigazítsuk. Egyelőre az látszik világosan, hogy a közmegegyezés szerint ugyanazon FNy szöveg különböző CNy fordításai között vannak „ekvivalensebbek" és „kevésbé ekvivalensek". Ennek az egyszerűsítő szemléltetése az egyenes szakasz, amelynek kezdőpontja a FNy szöveg, a végpontja pedig az ideálisan ekvivalens CNy szöveg, a létező CNy fordítások pedig e kettő között helyezkednek el:



3.3. ábra: A fordítási kontinuum

A javításkorpusz tehát biztosan modellje a fordításnak és a fordítás javításának. Ekvivalenciamodellé úgy válik, hogy lehetővé teszi legalább két a forrásszöveggel különböző szinten ekvivalens CNy szöveg szisztematikus összevetését. Ez azáltal lehetséges, hogy tudjuk: a publikált CNy szöveg az első CNy szöveg átalakításával jött létre, és joggal feltételezhetjük, hogy a két CNy szöveg hasonlít egymáshoz annyira, hogy különbségükből következtethetünk a javítási műveletekre, vagyis látjuk a javítás folyamatát. Ennek során két peremfeltételt (megszorítást) kell elfogadnunk:

- Feltételezzük, hogy a javítás során létrejövő CNy szöveg közelebb kerül az ideális CNy normához.
- Feltételezzük, hogy a javítás során az eredeti CNy szöveg transzformációja, nem pedig elvetése és teljes újraírása történik. Ez magasabb szinten annak feltételezése, hogy az első CNy szöveg előállítója rendelkezik a FNy szöveg lefordításának kompetenciájával.

Ha a javításkorpuszt valódi szerkesztőségi, illetve fordítóirodai folyamatok eredményeként létrehozott szövegekből állítjuk össze, a második feltételezést tapasztalati alapon megalapozottnak kell tekintenünk, ugyanis a fordítással foglalkozó szervezetektől a gazdasági realitás megköveteli, hogy előzetesen meggyőződjön a fordítók kompetenciájáról.

Hangsúlyozom, hogy a fentiekben csak modellt írtam le, új elmélet nem született. A javítási folyamat rekonstrukciójával és elemzésével ugyanakkor lehetővé válik a létező elméletek igazolása vagy cáfolata, illetve – ha szükséges – új elméletek létrehozása.

A korpusz empirikus vizsgálatával már kezdetben sikerült osztályoznom a javítás elemi műveleteit:

Művelet	Példa
explicitáció	more secure → biztonságosabb → a korábbiaknál biztonságosabb
terminológiai váltás	IT support → IT támogatással → informatikusi támogatással the Product business object → a termék üzleti objektum → a Product üzleti objektum upgrades → bővítéseket → frissítéseket
egyszerűsítés	to locate the information you want → a szükséges információ megtalálásához → keresheti meg a szükséges adatokat Use → Használja → ∅ via a background operation → háttérműveleten keresztül → a háttérben to allow for very long running applications → annak érdekében, hogy … rendelkezzünk → hogy az alkalmazások hosszú távon használhatóak maradjanak.
törlés	Use → Használja → ∅ via a background operation → háttérműveleten keresztül → a háttérben to allow for very long running applications → annak érdekében, hogy … rendelkezzünk → hogy az alkalmazások hosszú távon használhatóak maradjanak.
kiemelés	Using the … dialog, → a Data Source … használatával → Az űrlap használatának eszközei: for displaying the Product business object → a termék üzleti objektum adatainak megjelenítésére → amely a Product üzleti objektum adatait jeleníti meg.
beolvasztás	a Form with numerous controls → űrlapot, rajta több vezérlőelemmel → több vezérlőelemet tartalmazó űrlapot
Szubjektív javítás	a Form with numerous controls → űrlapot, rajta több vezérlőelemmel → több vezérlőelemet tartalmazó űrlapot

3.1. táblázat: A javítási műveletek osztályozása

Ezek azonban nem elemi műveletek abból a szempontból, hogy a számítógép milyen szerkesztési műveletek automatikus azonosítására készíthető fel, de a

géppel megtalált átalakítási műveletek elemzésével ezek a műveletek is kikövetkeztethetők.

Az ötlet nem teljesen új: a gépi fordítás iránti igény életre hívta azt a munkafolyamatot is, amelynek során a gépi kimenet emberi javításon esik át. Ennek lehetőségét már az ALPAC-jelentés is említi, és az utószerkesztés automatikus végrehajtásával kapcsolatban is folytak már kutatások (pl. Isabelle et al. 2004, Kranias et al. 2004). Nem tudok azonban olyan kutatásról, amely az emberi fordítás javításának gépi segítését célozza, a piacon elérhető néhány konzisztencia-ellenőrző eszköztől eltekintve.[20]

A javításkorpusz vizsgálata lehetőséget ad arra, hogy a konzisztencia-ellenőrző eszközök által végzett triviális ellenőrzési műveletek mellett lehetőség legyen a lektori javítások gépi tanulására és reprodukálására. Ennek lehetőségét a 4. fejezetben tárgyalom, a javításkorpusz részletes ismertetése során.

3.2. A fordítás új körülményei – a fordítástechnológia keletkezése

A következőkben a fordítástechnológiát mint a fordítástudomány kutatásának tárgyát mutatom be. Ennek során kimutatom a fordítástechnológia mikro- és makrostratégiájának hatását a fordítás folyamatára, különös tekintettel azokra a fordítástechnológiai folyamatokra, amelyek a fordítás megváltozott körülményeinek ellensúlyozására jöttek létre.

Közhely már, hogy a fordítás körülményei megváltoztak, ezért a fordítási feladatok jelentős része nem végezhető el a „hagyományos" elszigetelt alkotómunkával. Az elmúlt egy-két évtizedben többszintű folyamat játszódott le, ennek eredménye az a helyzet, amellyel a fordítóknak jelenleg szembe kell nézniük. Ez három alapvető tényezőt jelent:

(1) A teljes fordítási feladatot nem tudja egy fordító elvégezni a jelenleg szokásos határidőkkel.

(2) Az egy fordítóra jutó fordítás elvégzésére is lényegesen kevesebb idő jut, mint korábban. Nem ritka, hogy a fordítással foglalkozó szervezetek nem napban, hanem órában és percben határozzák meg a határidőt, és nem rendkívüli az egy napnál rövidebb határidő sem. Emiatt a fordítóknak nincs lehetőségük saját munkájuk áttekintésére és javítására.

(3) A fordítással foglalkozó szervezetek és az egyéni fordítók egyformán a csökkenő fordítási díjak nyomása alatt vannak. A megrendelők előírják fordítómemória használatát, és a FNy szöveg fordítómemória-beli analízise után fizetnek (tehát az ismétlődések és a korábban lefordított anyagok után nem).

Ezek a tényezők együtt azt eredményezték, hogy a fordításnak a piacon elérhető minősége lényegesen romlik. Ez a saját fordítóirodai, szerkesztőségi megfigyeléseim alapján nyilvánvaló, de a 3.1. és a 4. fejezetben leírt javításkorpuszból igazolható is. A minőségromlás a következőkben nyilvánul meg:

(1) A több fordító között szétosztott fordítás eredményeképp kapott CNy szöveg nem lesz egységes, sem a terminológiát, sem a regisztert, sem a szöveg többi jellemzőjét tekintve.

(2) Az egy fordítóra nehezedő időbeli nyomás miatt a fordítás átváltási műveletei „sekélyebben" történnek, vagyis a fordítónak egyre ritkábban van lehetősége elérni az ekvivalens CNy szöveg előállításához szükséges ekvivalenciaszintet. Bár ez elhamarkodott magyarázatnak tűnik, Komisszarov elmélete jól magyarázza azt a jelenséget, amikor a fordítás „széttöredezik", a szórend a forrásnyelvéhez válik hasonlóvá – mindennek az az oka, hogy a fordító a korábbiakhoz képest rövidebb szövegegységeket tud csak áttekinteni. Példa a korpuszból:

'[...] a Form with numerous controls for displaying data the Product business object' →
**'[...] űrlapot, rajta több vezérlőelemmel a termék üzleti objektum adatainak megjelenítésére'*

(3) A mikrostratégiai eszközök – elsősorban a fordítómemóriák – a szegmentumokat többnyire környezetükből kiragadva tárolják, emellett pedig olyan fordításokat is felajánlanak, amelyek forrásszövege csak részben egyezik az aktuális lefordítandó szegmentuméval. Az időbeli nyomás miatt nemritkán előfordul, hogy a fordító revízió nélkül illeszti a CNy szövegbe az adatbázisból kapott fordítást: ez nehezen észrevehető, sokszor „gépszerű" fordítási hibákhoz vezet.

A fordítástechnológia kialakulása erre a helyzetre válasz. Előbb alakultak ki a mikrostratégiai elemek – azok az erőforrások, amelyek az egyes szegmentumok fordítását teszik gazdaságosabbá, később pedig a makrostratégiai folyamatok, amelyek a párhuzamosítás, az időbeli túlterhelés és a mikrostratégiai eszközök hatását ellensúlyozzák.

3.3. A fordítás mikrostratégiája

A mikrostratégia elemei az értekezés írásakor már adottságnak számítanak, az eszközök igénybe vétele megszokott dolog.

A fordítástechnológia mikrostratégiájának vizsgálatakor fel kell tételeznünk, hogy a fordító a CNy szöveget fordítómemóriát alkalmazó számítógépes fordítási környezetben szerkeszti meg. A számítógépes fordítási környezetet a korábbiakban jellemeztem. Itt elsősorban a fordítómemória-használat és a fordítástudomány kölcsönhatásával foglalkozom, ezen belül pedig a következő három kérdésre keresem a választ:

(1) A fordítómemória-használat mennyiben tekinthető az ekvivalenciaszintek elméletében leírt átváltási műveletek modelljének?

(2) Hogyan befolyásolja a fordítási folyamatot a fordítómemória-használat?

(3) Hogyan – milyen nyelvtechnológiai eljárásokkal – csökkenthetők a fordítómemória-használat negatív hatásai?

A fordítómemória-használat mint az átváltási műveletek modellje

Komisszarov (1990) felfogása szerint a fordítási folyamat egésze átváltási műveletek sorozata. Ez annak a transzformációnak az egyfajta analitikus leírása, amelyet korábban fekete dobozként kezeltünk, s Nida (1964) modelljét követve az analízis → belső reprezentáción végrehajtott átváltás → szintézis folyamataként fogtuk fel. Azonban az átváltási műveletek során a fordító nem vonatkoztat el mindig a konkrét forrásnyelvi szövegtől, hanem a forrásnyelvi és a célnyelvi elemek között megfeleltetéseket ismer fel, és ezeket konkretizálja átváltási műveletekkel. Mindezek során mégsem szabad figyelmen kívül hagyni a közlés és a környezet mentális reprezentációját, mivel a forrásnyelvi közlésnek mindenképpen van kontextusa és konnotációja, s az adekvát megfelelés felismerése csak ezek felhasználásával lehetséges. Ezért azt mondhatjuk, hogy az átváltási műveletek sorozatában a fordító sokszor nem jut el a teljes analízisig, és a műveleteket különböző nyelvi szinteken végzi el. Erre bizonyítékként szolgálnak azok az esetek, amikor a fordító „nem találja el" teljesen az átváltáshoz szükséges ekvivalenciaszintet, és a CNy szöveg, kisebb-nagyobb mértékben eltérve a célnyelvi normától és konvenciótól, bizonyos értelemben közel kerül a forrásnyelvi szöveghez.

A fordítómemória-használat nem tekinthető az átváltási műveletek modelljének. Az értekezés írásakor létező eszközök pusztán a transzformáció bemenetét és kimenetét jegyzik fel, illetve egyetlen manipulatív műveletre, a helyettesítésre alkalmasak, a terminus és a szegmentum (nagyjából a mondat) szintjén.

Gondolhatunk arra, hogy feljegyezzük a forrásnyelvi közléstől a célnyelvi realizációhoz vezető átváltási műveletek sorát. Ehhez azonban fel kell jegyeznünk vagy a CNy realizációból fel kell ismernünk minden elképzelhető átváltási műveletet, amely, ha nem is lehetetlen, mindenképpen munkaigényes folyamat. Mindezzel együtt nem lenne haszontalan egy átváltási műveleteket is tartalmazó korpusz felépítése.

Ha az átváltási műveletek sorozatát le is tudtuk írni, tudatában kell lennünk, hogy az effektív fordítói műveletnek csak kis részét jegyeztük fel, mivel nem írtuk le az átváltási művelet mögötti döntéseket, illetve az azokhoz vezető körülményeket és felismeréseket: a kontextust és konnotációt. Ha nem áll rendelkezésünkre más, csak a forrásszöveg – tehát nincs mentális világismeretünk, éppen úgy, ahogy egy számítógépnek sincs –, akkor hiába tudunk visszajátszani tetszőleges átváltásiművelet-sorozatot, mert nem rendelkezünk azzal a kompetenciával, hogy dönteni tudjunk az átváltási műveletekről.

Kitérő: a gépi fordítás mint az átváltási műveletek modellje

Bár ennek az értekezésnek a gépi fordítás nem elsődleges tárgya, érdemes feltárni az összefüggést az átváltási műveletek által alkotott modell és a gépi fordítás egyes stratégiái között.

A következőkben csak a szabályalapú gépi fordítással foglalkozom, mivel a statisztikai gépi fordítás eljárásai ezen a helyen nem relevánsak – utóbbiakra a 4. fejezetben visszatérek. Fontos hangsúlyozni, hogy kihagyásuk nem állásfoglalás arról, hogy a szabályalapú vagy a statisztikai módszerek magasabb rendűek-e. A szabályalapú rendszerekre viszont igaz, hogy a fordítás jól definiált szabályok alapján végrehajtott manipulációs (~átváltási) műveletek sorozata által alkotott transzformáció, s mint ilyen, analógnak tekinthető a fordítástudomány átváltásiművelet-fogalmával.

A gépi fordítás hagyománya szerint, a transzformáció absztrakciós szintje alapján háromféle fordítási stratégiát különböztethetünk meg (Prószéky 1989):

(1) közvetlen fordítás;

(2) közvetítőnyelves fordítás;

(3) transzfer fordítás.

1. A közvetlen fordítás kizárólag lexikai átváltásra épül, a forrásnyelvi szöveg szintaktikai elemzése kizárólag a lexikai többértelműség feloldására szolgál. E stratégia műveletei: a forrásnyelvi lexika behelyettesítése célnyelvi lexikával, és a lexikális egységek átrendezése az elvárt célnyelvi szórendnek megfelelően.

Ha ezt modellnek tekintjük, akkor a következő hiányosságokat állapíthatjuk meg:

- Csak lexikai és korlátozott mértékű szintaktikai átváltási műveletek vannak. Lexikai szinten behelyettesítés, elhagyás és korlátozott mértékben beszúrás lehetséges.
- Csak szó- és szószerkezet-szintű átváltási műveletek vannak.
- Az átváltási döntés alapját kizárólag a forrásnyelv és a célnyelv közötti szótár elemei, a többértelműségek feloldására szolgáló korlátozott méretű környezetiszabály-halmaz és a szórend mechanikus átváltási szabályai képezik. Kontextust és konnotációt ez a stratégia nem vesz figyelembe.

Műszaki szempontból e megközelítés hátránya, hogy túlságosan függ az alkalmazott nyelvpártól. Minden szabályt és szótári egységet teljesen újra kell építeni, ha a rendszert újabb forrás- vagy célnyelvre készítjük fel.

2. A közvetítőnyelves (interlingvális) fordítás bizonyos értelemben törekszik a Nida-féle fordítási modell megvalósítására (talán nem véletlenül, hiszen körülbelül egy időben keletkeztek), vagyis a forrásszöveg analízisé-

nek eredménye a forrásszegmentum közvetítőnyelvi ábrázolása, s ebből egy független folyamat szintetizálja a célnyelvi szegmentumot.

A közvetítőnyelves rendszerekben az analízis és a szintézis is rendszerint a szintaktikai elemzésre, illetve generálásra szorítkozik, bár később már (például a német VERBMOBIL projektben – vö. Görz et al. 1996) korlátozott témájú szövegek szemantikai elemzésére is vállalkoztak. Mivel azonban a legnagyobb hangsúly a szintaxison van, a közvetítőnyelves rendszerek rendkívül érzékenyek a többértelműségre, s ezért a legtöbbször nagyon sok fordítási alternatívát állítanak elő.

Ha közelebbről megvizsgáljuk ezt a modellt, észrevehetjük azt is, hogy az interlingvális átalakítás nem a fordítandó közlés absztrakt ábrázolását jelenti, hanem egy olyan belső formalizmust, amely egy mesterséges nyelvet valósít meg, kontextus és konnotáció nélkül. Ez pedig azt jelenti, hogy az interlingvális fordítás tulajdonképpen nem egy, hanem két fordítási művelet.

Az interlingvális gépi fordítás létrehozását különben sem az emberi fordítás modellezésének igénye motiválta, hanem az az egyszerű műszaki/gazdasági szempont, hogy egyszerűsítsék az újabb nyelvpárok bevezetését.

Amennyiben az interlingvális gépi fordítást tekintjük az átváltási műveletek modelljének, a modellben a következő hiányosságokat vehetjük észre:

- Kizárólag lexikai és szintaktikai átváltási műveletek vannak.
- Kizárólag szó- és szószerkezet-szintű átváltási műveletek vannak. (vö. Klaudy 1999)
- Nincs tágabb kontextus és konnotáció.

3. A transzfer fordítórendszerek a fenti két „szélsőség" között kaptak helyet. „Kifejlesztésével az volt a cél, hogy csökkentsék a közvetlen stratégia túlzott nyelvpár-függőségét, ugyanakkor kiküszöböljék a közvetítőnyelves rendszer általánossága miatt megjelenő túlgenerálást és melléfordítást. A transzfer stratégiában a forrásnyelv és a célnyelv önálló, egymástól független »mélyszerkezeti« reprezentációkkal rendelkezik, ezért a fordítás három lépésből áll: analízis, transzfer, szintézis. A szintaktikai elemzés ezekben a rendszerekben nem olyan mély, mint a közvetítőnyelves fordítások esetében, hiszen az ott tárolandó további információk egy részét a transzfer fázis viszi a rendszerbe." (Prószéky, 2005)

A transzfer módszer mint átváltásiművelet-modell ugyanazokat a hiányosságokat mutatja, mint az interlingvális megközelítés. Mindenhol kiemelhető a kontextus és a konnotáció figyelembe vételének hiánya, ami pedig az átváltási műveletek mögötti emberi döntések legfőbb motívuma. Külön hangsúlyozni kell, hogy mindez a szó és a mondatszint közötti szinteken sincs meg, ami azt jelenti, hogy a mondaton belüli kontextus is figyelmen kívül marad.

A fenti mindhárom módszerben megfigyelhető a számítógép korlátozásait „kiszolgáló" túlegyszerűsítő felfogás: minél kisebb és kevesebb féle alapegysé-

get felhasználni, s ezeket minél egyszerűbb és általánosabb szabályokkal összefogni. Így érthető, hogy minden esetben a szó az alapegység, és abból építenek sekélyebb vagy mélyebb mondatfákat.

4. Köztes modellek. A magyarországi kutatás – amely, hasonlóan a többi gépi fordítási projekthez, elsősorban a hatékony fordítórendszer létrehozását célozza, nem pedig az emberi fordítási folyamat modellezését – két ponton haladja meg a három hagyományos megközelítést:

(1) Szavak helyett alul- vagy inkább rugalmasan specifikált frazeológiai mintákra épül.

(2) A fordítást valóban (lexikai és szintaktikai) átváltási műveletekként írja le.

Szavak helyett alul- vagy inkább rugalmasan specifikált frazeológiai minták. Számos olyan felfogás van (pl. Kis Á. 2004, Moon 1998, és általában a szövegnyelvészeti hátterű kutatások), amely szerint a lexikális egység elsősorban nem a szó, hanem adott esetben két vagy több szó kollokációja. Megfordítva: minden olyan megközelítés, amely természetes egységnek tekinti a szót, súlyos nehézségekkel találkozik, amikor két vagy több szót újra „össze kell ragasztania", hogy értelmes egységet kapjon. Kis Ádám egyenesen arra jut, hogy a morféma mint jelentést hordozó entitás túlnyúlhat a szóhatáron: ekkor pedig a szó nem valódi egység és a szóköz nem valódi elválasztójel.

Megjegyezzük, hogy a szóköz használata, vagyis a szavak különírása, egyáltalán a szavak írása nagyon sok esetben önkényes, ortográfiai preskripció eredménye. Ennek a szöveg olvashatósága végett van létjogosultsága, de nem könnyíti meg a szöveg számítógépes tanulmányozását, ahol minden rendszer elsősorban szavakra bontja (tokenizálja) a feldolgozandó szöveget, mert a szóközt programmal egyszerűen feldolgozható elválasztójelnek tekinti.

A magyarországi kutatásban kifejlesztett ún. MetaMorpho-formalizmus leginkább az angol „fixed expression" (szó szerint: rögzült kifejezés) fogalomhoz (Moon 1998) áll közel. Rosamund Moon létrehoz egy taxonómiát az angol nyelv rögzült kifejezései számára, s ebben nagyon fontos tényező, hogy a kifejezés mely pontokon és milyen mértékben rögzült. Ezzel felállítható egy skála a produktív szókapcsolat (pl. *'vasúti átkelőhely'*) és a többszavas lexéma (*'dugába dől'*) között, ahol helyet kapnak az igevonzat-keretek, az idiomatikus kifejezések és az egyéb frazémák.

A MetaMorpho-formalizmus erénye éppen az, hogy az ábrázolt nyelvi elem egyes részeit rögzíteni tudja, másoknak pedig az elképzelhető legnagyobb szabadságfokot adja. Egyszerűsített példa az angol *'a bottle of wine'* ábrázolására:

```
NP = DET(def = INDEF) + N(lex = „bottle") + PREP(lex = „of") + NP(object_nature = LIQUID)      (1)
NP[object_nature = LIQUID, lex = N.lex] = N(lex = „wine")                                      (2)
```

A szintaktikai címkék magyarázata:

NP – főnévi csoport
N – főnév

PREP – elöljáró

Az (1) szabály az *'a bottle of* <folyadék>' mintát írja le. A *<folyadék>* [a szabályban: `NP(object_nature = LIQUID)`] helyére tetszőleges főnévi csoport behelyettesíthető, amely rendelkezik az `object_nature = LIQUID` jeggyel.

A (2) szabály voltaképpen szótári szabály, amely a 'wine' főnévből főnévi csoportot állítja elő, és a lemmája alapján hozzárendeli a megfelelő jegyet. Fontos, hogy a mai számítógép-kapacitások mellett nem jelent problémát az összes folyadék felsorolása hasonló mintákkal, s a MetaMorpho-formalizmus eleve feltételezi, hogy az elemzési mintákat (mert ezek klasszikus értelemben véve inkább minták, mint szabályok) tartalmazó „nyelvtan" több százezer, esetleg több millió mintából áll.

A fenti leírás másik erénye, hogy a rendszerbe tetszés szerint lehet jegyeket bevezetni: nincs különbség szintaktikai és szemantikai jegy között, így, ha a felhasználás szempontjából fontos, hogy a rendszer figyelembe vegye az egyes dolgok halmazállapotát, akkor erre a célra külön jegyet is alkalmazhatunk (mint ebben az esetben az `object_nature`, amely különben az implementált rendszernek ebben a formában nem része).

Ez a megközelítés lehetővé teszi a szó- és mondat-, illetve szószerkezetszint közötti szinten a kontextus felhasználását, s így a szónál nagyobb egységek lexikai átváltását. Ez a triviális művelet ugyanis korábban nem volt elérhető a gépifordító-rendszerek számára.

A fordítás leírása átváltási műveletekkel. A MetaMorpho-formalizmus a fordítást valóban (lexikai és szintaktikai) átváltási műveletekként írja le, amelyek a szintaktikai elemzés közben végbemennek – ez megfelel az ekvivalenciaszintek elmélete által proponált felfogásnak, amely nem feltételezi a fordítási folyamatban a forrásszegmentum teljes mélységű analízisét.

A fordítás mechanizmusa e formalizmusban arra épül, hogy minden elemzési (frazeológiai, szintaktikai) szinten kicseréljük az egyes nyelvi elemeket feltételezett fordításukra. Ez a csere a *<forrásminta, célminta>* párok felhasználására épül, azonban a célminta elemeit meg kell feleltetni a forrásminta elemeinek, mivel a célnyelvi oldalon nem jön létre a forrástól független elemzési fa (elemzési erdő).

Példa a fenti mintákhoz rendelt átváltási műveletekre:

```
EN.NP = DET(lex = „a") + N(lex = „bottle") + PREP(lex = „of") + NP(object_nature = LIQUID)
HU.NP = DET[lex = „egy"] + N[lex = „üveg"] + NP                                          (3)
```

A 'bor' pedig így kerül a fordításba:

```
EN.NP[object_nature = LIQUID, lex = N.lex] = N(lex = „wine")
HU.NP = N[lex = „bor"]                                                                   (4)
```

Látható, hogy a (3) szabályban egyes, szintaktikai címkékkel jelölt szimbólumok meg vannak egymásnak feleltetve. A megfeleltetés alapja a szintaktikai címkék

egyezése. Ha a forrásoldalon több egyforma szintaktikai címkével jelölt szimbólum van, akkor sorszámokat kell alkalmazni. E megfeleltetés ismeretében írhatók le a helyettesítési szabály hatókörén belüli átváltási műveletek: a kihagyás, a beszúrás és az áthelyezés. A jelen példában kihagyást látunk: az elöljáró kimarad a magyar fordításból.

Összefoglalva: a gépi fordítás módszereit bizonyos szempontból sokkal inkább lehet az átváltási műveletek modelljének tekinteni, mint a gépi fordítástámogatás műveleteit. Az előbbi módszerei között pedig van olyan, amely közvetlenül is alkalmas az átváltási műveletek produktív modellezésére. A gépi fordítástámogatás alkalmazásakor azonban megvan az ember lehetősége az átváltások tényleges végrehajtására, az előbbiben azonban a döntést az automatára ruházzuk. Lényeges különbség még, hogy a gépi fordítástámogatás eszközei az emberi fordító átváltási döntéseit játsszák vissza, míg a gépifordítórendszerek sajátos algoritmikus döntéseik alapján kombinálják az egyes műveleteket – ezért tűnnek alkalmasint rosszabb nyelvi minőségűnek, illetve a célnyelv struktúráitól és lexikájától távolibbnak a géppel automatikusan előállított fordítások.

A fordítómemória-használat hatása a fordítás folyamatára

Sem a jelenlegi, sem a kutatási szakaszban levő fordítómemória-eszközök nem próbálják megismételni az ember által végezhető átváltási műveleteket, hanem – megfelelően felismert forrásszöveg esetén – az átváltási műveletek eredményét kínálják fel. Ez az ember számára – a felszínen – azt jelenti, hogy a gép mégis elvégezte helyette az átváltási műveletek egy részét, másokat pedig rábízott (hogy melyeket igen és melyeket nem, az ezen a ponton nem jósolható meg).

A fordítómemória-használat az átváltási műveletek szempontjából azt jelenti, hogy a számítógép visszajátszik egyes tárolt átváltásiművelet-sorozatokat, vagyis felkínálja a korábban már tárolt forrásszegmentumhoz tartozó egyik lehetséges fordítást, amelyet a fordítónak pedig ki kell igazítania. Feltételezhetjük, hogy ez némiképp másfajta kompetenciát igényel, mint a „tiszta" fordítás, mivel itt egyes átváltási műveleteket vissza is kell vonni. Ez egyfajta lektorálás, ám a fordításjavítás jóindulatú előfeltevése, miszerint a kapott fordítás az aktuális forrásnyelvi szegmentum ekvivalense, nem tartható. A fordítómemória-használat során a legtöbbször olyan segítséget kapunk az adatbázistól, ahol

- az adatbázisban tárolt FNy szegmentum csak részben egyezik az aktuális FNy szegmentummal;
- az adatbázisban tárolt FNy szegmentum egyezik az aktuális FNy szegmentummal, de az tárolt FNy szegmentum eredeti környezete eltér az aktuális FNy szegmentumétól.

Ezért a felajánlott CNy fordítás nem szándékolt ekvivalense az aktuális FNy szegmentumnak, így a fordítómemória-találatok kiigazítása nem egyszerű lektori művelet.

A fordítómemória-használat negatív hatásainak csökkentése

A fordítástechnológia mikrostratégiájának kutatásában egyelőre nincs elmozdulás attól a ponttól, hogy a rendszerek egyelőre csak helyettesítésre képesek, méghozzá lexikai és mondatszinten. Az utóbbi a fordítómemóriák működésének eredménye, de lexikainak tekinthető abban az értelemben, hogy rögzített nyelvi jelet cserél ki rögzített nyelvi jelre.

A fordítómemóriák továbbfejlesztésével kapcsolatos kutatásnak két iránya van:

- Hatékonyságnövelés: a fordítómemória-találatok arányának és gyakoriságának növelése új hasonlósági keresési módszerek bevezetése által. Itt helyet kap a nyelvi elemzés alkalmazása és a jelenlegi módszerekkel irrelevánsnak tekintett tárolt szegmentumok építőelemként való felhasználása is (Kis et al. 2004, Callison-Burch 2005). Utóbbiakat az értekezés írása idején a leghatékonyabban az ATRIL cég által kifejlesztett Déjà Vu rendszer alkalmazza.
- Utószerkesztés: algoritmusok kifejlesztése az adatbázisból kiemelt CNy fordítás módosítására úgy, hogy az az aktuális FNy szegmentum ekvivalensévé váljon. Ez a törekvés a gépi fordítás kutatásában is megjelenik. (Vö. Isabelle et al. 2007, Kranias et al. 2004, Hodász G. et al. 2004).

E kutatások irodalmi forrásai viszonylag gyérek, mivel az alkalmazott algoritmusok kereskedelmi termékek védett és üzleti titokként kezelt részeiként vannak megvalósítva.

A 4. fejezetben vázlatosan bemutatok két eljárást, amelyek arra irányulnak, hogy a fordítómemóriák és a terminológia együttes felhasználásával bizonyos esetekben hatékony, részben már utószerkesztett fordítási javaslatot adjon. Ez ugyan nem garantálja a javasolt CNy szövegnek az aktuális FNy szegmentummal való teljes ekvivalenciáját, de a szándékolt ekvivalenciát igen. Másképp fogalmazva: helyreállítja a jóhiszeműségi hipotézist, mely szerint a javasolt CNy szöveget az aktuális FNy szegmentum fordításának *szánták,* s így az adaptálási feladatot a hagyományos lektori művelethez közelíti.

3.4. A fordítástechnológia makrostratégiája

A fordítástechnológia makrostratégiája a fordítás mikrostratégiai műveleteit rendezi jól definiált folyamatba. Ennek megfelelően három rendeltetése van:

- Az új típusú (megnövekedett volumenű és szorosabb határidejű) fordítási feladatok elvégzésének biztosítása szervezéssel, munkafolyamat felállításával; a projekt költségvetésének meghatározása és ellenőrzése;

- A fordítás műszaki szinergiájának biztosítása. A fordítás nem öncélú, minden esetben valamilyen műszaki termék – könyv, weboldal, szoftver stb. – előállításához kapcsolódik. Az összetett fordítási projektek magukban foglalják a FNy szöveg műszaki előkészítését és a kész CNy szöveg műszaki előállítását is;
- Minőségbiztosítás: a volumen-idő nyomás negatív hatásainak enyhítése. Ez az utólagos ellenőrzésen túl a fordítást megelőző és a fordítás közben érvényesülő minőségbiztosítási intézkedéseket is jelent. A legjobban kidolgozott minőségbiztosítási módszerek a terminológiakezelésben működnek; ezeket az 5. fejezet (az 5.1. rész) részletesen ismerteti.

A következőkben vázlatosan bemutatom a makrostratégia elemeit, majd két olyan makrostratégiai részfolyamatot, amely széttagolt projektekben, illetve kiélezett időbeosztás esetén is lehetővé teszi a fordítás minőségbiztosítását. Az utóbbiak a részben saját fejlesztésű MemoQ rendszer részeként évek óta segítenek fordítási projekteket.

A makrostratégia elemei

A fordítás mindig nagyobb folyamat része. Ez azt jelenti, hogy amennyiben a fordítás hatékonyságát és folyamatát vizsgáljuk, sohasem vonatkoztathatunk el a fordítási projekt céljától, illetve körülményeitől (Lengyel et al. 2004, Lengyel 2006).

Az összes lehetséges makrostratégia bemutatása túlhalad az értekezés keretein, ezért a legegyszerűbb egy esettanulmányt ismertetni. A példa egy könyvkiadói fordítási projekt, amelynek célja egy angol eredeti nyelvű szakmai (informatikai) kiadvány megjelentetése adott határidőre. A könyvkiadó megszabja a fordítás körülményeit, mert a könyvkiadásnak meghatározott technológiája van, amelyben a fordítást el kell helyezni. A kontextusból kiemelt fordítási feladat az összefüggő dokumentum rövid határidejű lefordítása, ami lényegében a fordítás párhuzamosítását jelenti.

A példabeli munka terjedelme 151 738 szó, 784 028 karakter, kb. 600 oldal. A rövid határidő azt jelenti, hogy a teljes könyvkiadási folyamat végigviteléhez (a forrásszöveg kézhezvételétől a nyomdából való kiszállításig) 10-12 hét áll rendelkezésre. Egy informatikai szakkönyvkiadó, ha naprakész fordításokat akar megjelentetni, nem alkalmazhat ennél hosszabb átfutási időt.

Ha a fordítási folyamatot párhuzamosítással akarjuk felgyorsítani, a munkát több fordító egyidejű foglalkoztatásával kell megoldanunk. Ekkor azonban nagy hangsúllyal merül fel a konzisztencia kérdése, így összességében a következő minőségi problémákra kell megoldást találnunk a munkafolyamatban:

(1) Teljesség: a rövid határidővel készülő fordítások tipikus hibája, hogy mondatok, bekezdések kimaradnak. Ezeket fel kell ismerni, és pótolni kell.

(2) K o n z i s z t e n c i a : a szóhasználat és a stílus egysége(ssége) a teljes szövegben. Ebben egyetlen fordító alkalmazása esetén is akadnak hibák, amelyek azonban, ha a könyv különböző fejezeteit más és más fordítja, elkerülhetetlenek. Kezdetnek elegendő kétféle konzisztencia megkülönböztetése:

   a) T e r m i n o l ó g i a i k o n z i s z t e n c i a : a szakszöveg vázát alkotó terminológia egységes és helyes fordítása a teljes szövegben. Ez a terminológia teljességét és egységességét egyaránt jelenti, ahol az előbbi megköveteli, hogy minden, az eredeti szövegben terminológiai szándékkal leírt kifejezést a terminológia részeként kezeljünk.

   b) F r a z e o l ó g i a i k o n z i s z t e n c i a : az összekötő (a diskurzust felépítő) elemek egységes fordítása, vagyis egységes regiszter és egységes stílus.

A fentiekben csak a párhuzamosításból eredő minőségi problémákat vázoltam fel, és nem foglalkoztam a fordítás minőségének általános kérdéseivel, amelyeket Dróth Júlia különben kimerítően elemez (Dróth 2002). Az értekezésben általánosan is igyekszem kerülni a preskriptív megközelítést, és a minőségbiztosítást magát is tudományos vizsgálódás tárgyává kívánom tenni. Ennek megfelelően itt nem az általam kívánatosnak tartott folyamatot, hanem a szerkesztőségi munka során kialakított, illetve tapasztalt folyamatot írom le.

A fordítás minőségének biztosítását általában utólagos javítással valósítják meg. Ez azonban bizonyítottan a legköltségesebb módja a minőség biztosításának (vö. Lengyel 2006), ezért az ilyen projektek esetén mindig meg kell könnyíteni előkészítéssel. Az utólagos ellenőrzés a könyvkiadói projekt esetén három lépést jelent:

(1) szakmai ellenőrzés (lektorálás vagy kontrollszerkesztés);

(2) nyelvi ellenőrzés (olvasószerkesztés);

(3) tipográfiai ellenőrzés (korrektúra, esetleg két korrektúraforduló).

Az egységesség biztosítása azonban megköveteli, hogy az egyes minőségbiztosítási lépéseket egy-egy személy végezze, vagyis ezek a lépések önmagukban nem párhuzamosíthatók. Ha viszont sok a hiba, akkor különösen az (1) és (2) minőségbiztosítási lépéssel könnyen elveszíthetjük azt a megtakarítást, amelyet magának a fordításnak a párhuzamosításával elértünk.

Ezért makrostratégia a fordítás előkészítésébe legalább annyi energiát fektet, mint az utólagos minőségellenőrzésbe. Ez – egyelőre dióhéjban – a következő feladatokat jelenti:

(1) Közös fordítási erőforrások biztosítása: ez azt jelenti, hogy a későbbiekben részletezett számítógépes fordítási erőforrásokhoz, benne a terminológiához, minden fordító egyformán és naprakészen hozzáfér.

(2) A terminológia előzetes kialakítása (a példában 1700 tétel): az így kialakított terminológiai szószedetben leírt célnyelvi megfelelőktől a fordítók nem térhetnek el, akkor sem, ha nem értenek vele egyet. A terminológiának ebben

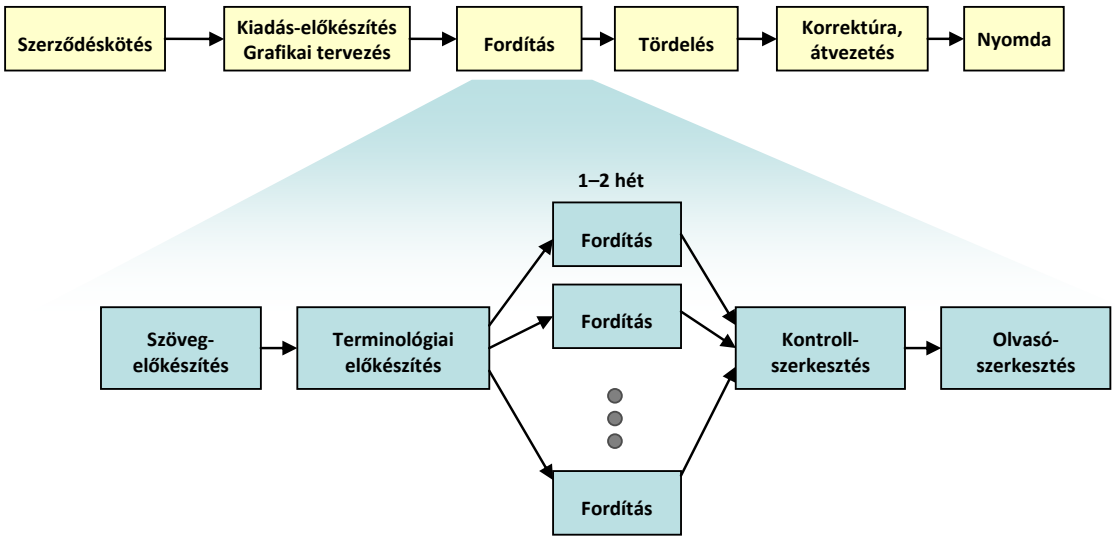
az esetben is lehetséges valamiféle nemlineáris fejlesztési protokollja, azonban ezzel nem foglalkozunk az előadásban.

(3) Előzetes stilisztikai/frazeológiai útmutató kiadása: 6-12 olyan fordítási utasítás kiadása, amelyben meghatározzuk a forrásszövegre jellemző, abban gyakran előforduló frazeológiai, szövegszervezési (diskurzusbeli) fordulatok fordítását. Példa: a *'use the Enter key to confirm your translation'* – a szövegben ismétlődő – típusú kifejezésben a *'use'* ige nem fordítható, a kifejezés helyes fordítása: *'a fordítást az Enter billentyűvel erősítsük meg'*.

(4) Elosztott fordítási infrastruktúra kialakítása: olyan számítógépes, hálózatos környezet kialakítása, amelyben a fordítási erőforrások – köztük a terminológia – folyamatosan és naprakészen hozzáférhetők a fordítók számára.

(5) Kommunikációs infrastruktúra kialakítása: a fordítás párhuzamosítása miatt a forrásszöveg több dokumentum formájában halad végig a munkafolyamaton. A kommunikációs infrastruktúra elsődleges célja a dokumentumok útjának jól definiált biztosítása, nem pedig a fordítók közötti kommunikáció hatékony biztosítása (tudniillik az e-mail és a csevegőprogram triviális módon rendelkezésre áll, ezek kialakítására nem szükséges külön – fordításspecifikus – erőfeszítés).

A munkafolyamatot az alábbi ábra szemlélteti a legjobban:



3.4. ábra: A könyvfordítási munkafolyamat egyszerűsített ábrája

A felső – lineáris – diagramon a könyvkiadás általános folyamata látható. Ezt itt nem részletezzük. Kiemeljük azonban a szöveggel foglalkozó részt, amely a felső diagramon a „Fordítás" cím alatt szerepel. Valójában ez összetett folyamat; lépései a következők:

(1) Szöveg-előkészítés: a forrásszöveg formátumának és kódolásának technikai átalakítása, hogy a rendelkezésre álló informatikai eszközökkel lehetséges legyen a terminológiai előkészítés és a fordítás.

(2) Terminológiai előkészítés: a lehető legteljesebb terminológiai szószedet előállítása a forrásszövegből, amelynek célnyelvi megfelelői kötelezően alkalmazandók a fordítás során. Ekkor történik a fordítási útmutató kialakítása is. Ezzel a későbbiekben részletesen foglalkozunk.

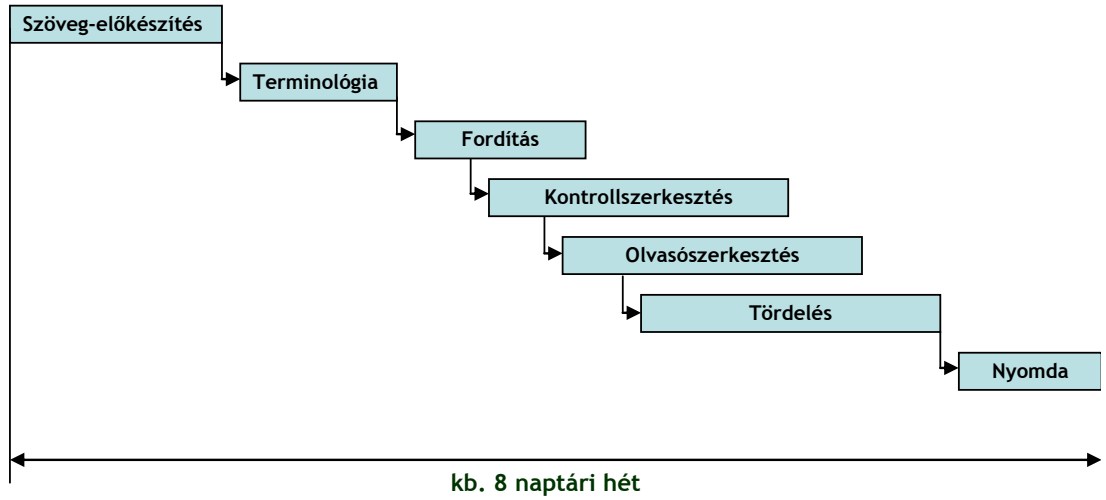
(3) Fordítás: a párhuzamosított fordítási lépés: a kötet fejezeteit gyakorlatilag egy időben fordítják le. A hagyományos fordítási munkától eltérést jelent a számítógépes fordítási erőforrások kiterjedt – hálózati – használata.

(4) Kontrollszerkesztés: a fordítás szakmai helyességének, terminológiai konzisztenciájának, illetve az előírások betartásának ellenőrzése. Kétnyelvű párhuzamos szöveg alapján történik.

(5) Olvasószerkesztés: a fordítás nyelvi helyességének ellenőrzése.

A fenti egyszerűsített munkafolyamatban nem lehetséges a visszacsatolás a fordítók felé. Az olvasószerkesztés (4) és a kontrollszerkesztés (5) rövid kérdések formájában kerülhet ugyan visszacsatolásos kapcsolatba (ezt az ábra nem jelzi), de a határidők jellemző rövidsége miatt egyébként is célszerű minimalizálni a visszacsatolást.

A folyamat kb. 8 naptári hét alatt vihető végig; ezt az alábbi, Gantt-diagramra emlékeztető ábrával szemléltethetjük:



3.5. ábra: A fordítási munkafolyamat időbeli lefolyása

A fordítás párhuzamosítása mellett, kihasználva, hogy a forrásszöveget részdokumentumokra bontottuk, a minőségellenőrzési folyamat elemei is átfedhetnek magával a fordítással, illetve egymással. Az egyes fázisok jellemzően egy hét eltolással kezdődnek, a fordítás teljes ideje 10-14 nap.

A következőkben – a fenti felosztás alapján – sorra veszem a makrostratégia minőségbiztosítási elemeit, vagyis azokat, amelyek a párhuzamosítás és a rövid határidő okozta nyomást hivatottak ellensúlyozni. Előbb azonban teszek egy kitérőt, amelynek során felvázolom a fordítás minősége és a fordítástechnológia – azon belül leginkább a számítógép – közötti kapcsolatot.

A fordítás minőségbiztosítása és a számítógép

Bár mindvégig hangsúlyozom, hogy a fordítástechnológia nem kizárólag számítógépes eszközök alkalmazását jelenti, a fordítás mint gazdasági tevékenység szervezése során a számítógépes eszközöknek – feltételezett nyelvi képességeikhez viszonyítva is – igen nagy szerep jut a minőségbiztosításban. Ezzel kapcsolatban kézenfekvő gondolat volna, hogy a számítógéppel elsősorban a fordítás javítását kellene automatizálni. Ezt azonban a gazdasági gondolkodás legalábbis kérdésessé teszi. A következőkben igyekszem megmutatni, hogy melyek a számítógép lehetőségei és korlátai a fordítás javításában, majd sorra veszem a minőségbiztosításban ténylegesen alkalmazott technológiai elemeket.

A jelenleg létező makrostratégiák rendkívül határozottak a minőség kérdésében: jól definiált eszközöket alkalmaznak a fordítási „hibák" észlelésére és javítására. Nem is tehetnek másként, mert a fordítás gazdasági tevékenység, amelynek költségei és kockázatai vannak, s ezek jó tervezése és ellenőrzése végett van szükség a minél szigorúbb technológiai fegyelemre.

A kérdést elméleti szempontból vizsgálva azonban emlékeznünk kell arra, hogy nem rendelkezünk kísérletileg igazolt ekvivalenciakritériumokkal, mégis megfogalmazzuk a fordítás minőségi kritériumait. Ezek leginkább szubjektív tapasztalatok rendszerbe foglalt leírásai, amelyek alapján kialakult valamiféle közmegegyezés. Közülük is talán a legteljesebb Dróth (2002) leírása. Azonban ezeket semmiféleképpen sem szabad leértékelni, mert a kísérletileg bizonyított elmélet híján is szükségünk van olyan szempontrendszerre, amelyek lehetővé teszik a fordítások értékelését és javítását mind a fordítási projektekben, mind pedig a fordítás oktatásában.

A rendszerszerű, taxonomikus leírások gyakorlati értéke kimondottan nagy, mert megkönnyítik a számítógép alkalmazását a feltételezett fordítási hibák felismerésében és javításában. Kis Ádám (1997) a gépi helyesírás-ellenőrzés szempontjából hangsúlyozza az írással kapcsolatos normák szisztematikus, „gépszerű" voltát.

Dróth (2002) Komisszarov ekvivalenciaszint-elméletéhez alkalmazkodva a különböző nyelvi szinteket veszi sorra a szempontrendszer kialakításában:

a) Szövegen kívüli tényezők, kommunikációs helyzet
b) Szövegszint:
   - Döntések: retorikai cél, műfaj, regiszter – szaknyelv
   - Kohézió
   - A tagmondatok és a mondatrészek logikai és tematikus sorrendje
c) Szintaktika
d) Lexika, terminológia
e) Felszíni elemek: helyesírás, szövegszerkesztés stb.

Hangsúlyozni kell, hogy a számítógép nélkülözi a világismeret autonóm ábrázolásának és felhasználásának képességét, ezért magát a világismeretet is.

A szöveggel kizárólag mint karaktersorozattal találkozik, s bár különböző eljárásokkal a szöveghez rengeteg – lexikai, szintaktikai, szemantikai – adatot lehet hozzárendelni, mindenhez a szöveg felszíni jelenségeinek vizsgálatával jutunk el. A számítógép nem tud hozzáférni a konkrét kommunikációs helyzethez, a retorikai célhoz, a műfajhoz, bár a felszíni jegyek alapján ezekről közelítő információhoz juthat. Ennek alapján vizsgáljuk meg, hogy a számítógép milyen szolgáltatásokat nyújthat az egyes nyelvi szinteken (a betűk a fenti felsorolásnak felelnek meg):

a) A szövegen kívüli tényezők nem hozzáférhetők, ezért ezek esetében semmilyen számítógépes szolgáltatás nem értelmezhető.

b) A szövegszinten – az eszközök mai fejlettsége mellett – lehetséges a regiszter felismerése és értékelése. Meghatározott regisztert jelezhetnek a szöveg egyes, nem feltétlenül tartalmas szavai, szókapcsolatai és morfoszintaktikai jellemzői. Ugyanazon regiszter pedig kizár meghatározott szavakat, szókapcsolatokat és morfoszintaktikai jellemzőket. Ezek számítógéppel felismerhetők, és bár nem folytattak eddig ilyen kutatást, a módszer egyszerű eszközökkel megvalósítható.

c) A szintaktikai szinttel paradox módon nem sokat tud kezdeni a számítógép. Itt ugyanis azt kell megállapítani, hogy a fordító megfelelően ültette-e át a forrásnyelv egyes szintaktikai szerkezeteit, illetve a célnyelv normájának megfelelő szintaktikai szerkezetek jöttek-e létre.

A szintaktikai szerkezetek félreértelmezését csak a CNy szöveg szintaktikai elemzésével lehet felismerni, ahol elméletben olyan nyelvi jelenségeket kell keresni, amelyek jellemzőek az adott forrásnyelv egyes szintaktikai szerkezeteinek félrefordításában. Nem beszélve arról, hogy – legalábbis a magyarban – eddig nem találkoztam ilyen jelenségek szisztematikus formális gyűjtésével. A feladat azért oldható meg nehezen, mert a szintaktikai elemzés rendszeresen áldozatául esik a többértelműségeknek és a rosszul kiválasztott értelmezés okozta téves összevonásoknak. Ez pedig mindaddig így lesz, amíg nem lesz természetes, hogy a szintaktikai elemzés során a számítógép legalább a kontextushoz hozzáfér. E nélkül ugyanis a szintaktikai félrefordítások ellenőrzése nem valósítható meg megbízható módon.

Ami a nyelvi normának való megfelelést illeti: szövegszerkesztőkben – legalábbis egyes nyelvekhez – rendelkezésre állnak nyelvhelyesség-ellenőrző programok, amelyek a szöveg egyes, szóhatáron túli lehetséges hibáit ismerik fel. Ezek azonban valamennyien feltételezik, hogy eredetileg is a célnyelven írt szöveg ellenőrzésére használják őket. Ha pedig elfogadjuk a fordítási folyamat dinamikus-párhuzamos értelmezését, vagyis azt, hogy a fordító „nem mindig vonatkoztat el teljesen a forrásnyelvi formától" (Klaudy 1999), feltételeznünk kell, hogy a fordítás során létrehozott CNy szöveg sajátos szintaktikai jellegzetességeket, esetleg hibákat mutat, amelyeket alapvetően befolyásolnak a FNy szöveg szintaktikai jellemzői. Kézenfekvő lenne tehát olyan nyelvhelyesség-ellenőrző programokat készíteni, amelyek a fordításnyelv jellemző hibáira van-

nak felkészítve, és ezeket próbálják a CNy normára javítani. Ez olyan nyelvhelyesség-ellenőrző programokban valósítható meg, amelyek nem a jól, hanem a rosszul formált nyelvtani szerkezeteket próbálják felismerni, vagyis nyelvmodell helyett hibamodellt alkalmaznak. Ilyen például a Helyesebb nevű közismert magyar nyelvhelyesség-ellenőrző modul. Ez viszont – figyelembe véve a fejlesztés munkaigényét és a forrásnyelvek számát, illetve az adott forrásnyelvekkel foglalkozó fordítók populációját – kevés kivételtől eltekintve nem tekinthető gazdaságilag reálisnak.

Elfogadható alternatívát jelent viszont annak felismerése, hogy a fordítók – a forrásnyelvtől függően – hajlamosak visszatérő, tipikus hibákat elkövetni a CNy szövegben. E hibákat a fordítás javítása során rendszerint ugyanolyan vagy hasonló módon javítják. A javítások szisztematikus észlelése, elemzése és reprodukciója pedig elvezet az automatikus–félautomatikus fordításjavító programok kifejlesztéséhez. Ennek alapját szolgálja a SZAK javításkorpusz, amelynek felépítését és felhasználását a 3.1. és a 4. fejezetben ismertetem.

d) A lexika, de leginkább a terminológia ellenőrzésében nagy szerep juthat a számítógépnek. A terminológiai konzisztencia és a terminológiahasználat egyszerűen ellenőrizhető, ha rendelkezésre áll megfelelő terminológiai gyűjtés. A számítógép azonban nem tudja olyan terminológiai elemek használatát ellenőrizni, amelyek nem állnak rendelkezésre a terminológiai adatbázisban.

e) A felszíni jegyek ellenőrzése viszonylag egyszerű, amennyiben a helyesírás ellenőrzése alatt a szövegszerkesztők szokásos helyesírás-ellenőrző funkcióját értjük. Ha nem, akkor a *c)* nyelvi szinthez tartozó, nyelvhelyesség-ellenőrzésről írott gondolatok az érvényesek.

A fordítás szempontjából érdekes, hogy a helyesírás- és a nyelvhelyesség-ellenőrző programok jelenleg nem veszik figyelembe a fordító anyanyelvét. A hibákat általában valamiféle előzetes hibastatisztika alapján vagy spekulatív módszerekkel javítják. A fordítás, illetve általában a szöveg-előállítás sokat nyerne az olyan helyesírás-ellenőrző programoktól, amelyek „tanulnak" a felhasználójuk által „elkövetett" és javított hibákból.

Hangsúlyozom, hogy a fenti szolgáltatások nem teszik képessé a számítógépet arra, hogy a fordításokat értékeljék, hiszen csak egyes, a fordítás minőségével kisebb-nagyobb korrelációban álló felszíni és jellemzően mennyiségi jellemzőket állapítanak meg. A fordításokat továbbra is az ember értékeli, de a gépi szolgáltatások fontos adatokkal szolgálhatnak a fordítások kijavításához.

A makrostratégia minőségbiztosítási elemei

Egyszerűsítve a korábbi leírásban olvasható felosztást, a makrostratégia három átfogó fázisból áll:

(1) Előkészítés
(2) Végrehajtás
(3) Utófeldolgozás

A minőségbiztosítás mindhárom fázisban megjelenik. Lengyel (2006) szerint a minőségbiztosítás az utó-feldolgozási fázisban a legköltségesebb, vagyis már gazdasági megfontolásokból sem célszerű mindent az utólagos ellenőrzési fázisokra hagyni. A következőkben mindhárom fázis esetén áttekintem az általánosan alkalmazott, illetve lehetséges minőségbiztosítási módokat, különös tekintettel az általam kidolgozott két részfolyamatra.

Előkészítés. Az előkészítés legfontosabb eleme a terminológiai előkészítés. Azonban a terminológia kezelése a teljes munkafolyamaton végighúzódik, ezért indokolt külön terminológiai rész-munkafolyamatról beszélni. A terminológiai munkafolyamattal részletesen is foglalkozom az 5. fejezetben, így erről itt nem esik több szó.

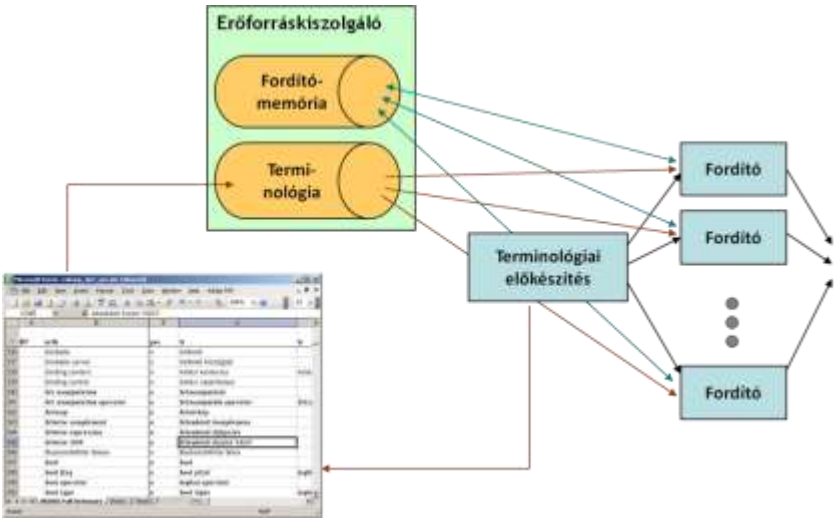
Az előkészítési fázisnak fontos eleme a technikai és a frazeológiai előkészítés. Előbbi a FNy szövegek olyan formára hozását jelenti, amely mellett a fordítók a lehető leglevesebb erőfeszítéssel képesek lesznek formatartó módon előállítani a CNy szöveget, mindezt anélkül, hogy a szövegformázás technikájára figyelniük kellene.

Utóbbi – a fordítási útmutató kialakítása mellett – az összetett FNy szöveganyag belső hivatkozásainak kezeléséhez szükséges. A fordítás megkezdése előtt fel kell mérni, hogy a FNy szövegben melyek azok az elemek, amelyekre hivatkozni lehet. Ilyenek a fejezetcímek, az ábraaláírások, műszaki vagy informatikai tárgyú szövegek esetén pedig a leírt műszaki eszköz kezelőfelületén megjelenő szövegelemek. Egy szakkönyv lefordítása például kezdődhet a tartalomjegyzék lefordításával.

A szoftver- és weboldal-honosítási folyamatokban a frazeológiai egységesség biztosításába általában beleértik a fordítómemória előzetes feltöltését is: ez meglehetősen mechanikus művelet, a szövegben egy bizonyos gyakoriság fölött előforduló szegmentumok előzetes lefordítását jelenti.

Végrehajtás. A fordítás konzisztenciája és minősége akkor biztosítható a legkevesebb erőfeszítéssel, ha a fordítók számára megfelelő infrastruktúrát alakítunk ki. Erre azért van szükség, mert így szabályozható, hogy a fordítók milyen erőforrásokhoz és kommunikációs eszközökhöz férnek hozzá, illetve ezek segítségével rövidíteni lehet a fordítási – terminológiai, frazeológiai, illetve általában az átváltási – problémák megoldásához szükséges kutatás idejét.

Ez triviálisan a mikrostratégiai erőforrások (fordítómemória és terminológiai adatbázis) alkalmazását jelenti. A lényeges változás a hagyományos (individuális) munkához képest az, hogy ezek az erőforrások közösek, a fordítók hálózatba kapcsolt fordítási környezetben dolgoznak:



3.6. ábra: A hálózatba kötött fordítási környezettel való munka sémája

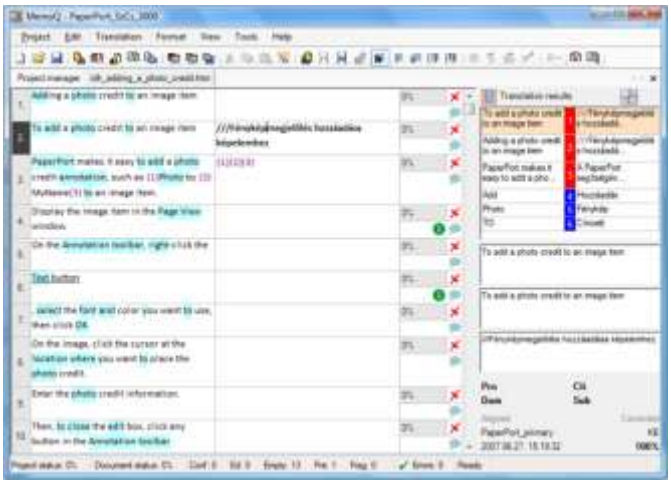
A 3.6. ábra jobb oldalán a fordítási munkafolyamat ábrájából (3.4. ábra) kivágott részlet látható. Minden fordító a saját számítógépén dolgozik. A gépeken olyan fordítástámogató eszköz fut, amellyel – az interneten át – elérhető az ábra közepén látható erőforrás-kiszolgáló. A felhasználók folyamatos kapcsolatban vannak a kiszolgálóval. (Feltételezzük, hogy mindenki otthon dolgozik, de széles sávú, rögzített díjas internet-hozzáférése van.) Ez azt jelenti, hogy amikor új szegmentumra (mondatra) lépnek, a rendszer automatikusan a kiszolgálóhoz fordul, és letölti az adott szegmentumhoz tartozó terminológiai szócikkeket, illetve lekérdezi a fordítómemóriát. Amikor pedig a felhasználó befejezi egy szegmentum (mondat) fordítását, az új fordítás automatikusan a kiszolgálóra kerül, és azonnal láthatóvá válik az összes többi fordító számára. Ugyanez történik azokkal a terminológiai szócikkekkel, amelyeket a fordítók eközben felvesznek az adatbázisba. Tapasztalatom szerint az utóbbira egy átlagos könyvfordítási projekt során 10-100 alkalommal kerül sor, vagyis az előkészített terminológiai anyag kevesebb mint 1%-át érinti.

A fordítási környezet implicit módon a következő szolgáltatásokkal segítheti a minőségbiztosítást (ezeket a MemoQ fordítási környezet meg is valósítja):

- Közös hálózati fordítómemória: amit egyvalaki lefordított, mindenki más azonnal látja, és ezért nem kísérli meg másképp lefordítani. Ez – bár a szegmentumok kevesebb mint 20%-át érinti – implicit módon hozzájárul a frazeológiai/stilisztikai konzisztencia biztosításához.
- A terminológia automatikus kijelölése: a fordító azonnal látja, hogy mit kell egységesen fordítani, és a CNy megfelelő is helyben rendelkezésére áll. Nincs szükség visszaemlékezésre és külön keresésre, ezért nagyobb a valószínűsége, hogy a fordító használja majd a normatív szószedetet, ezzel pedig növekszik a terminológiai konzisztencia. A több ezer tételes előkészített terminológiai adatbázis használatához az automatikus kijelölésre feltétlenül

szükség van, mert egyetlen fordító sem tudja megtanulni az összes FNy tételt, és észrevenni, hogy azok közül melyek szerepelnek az aktuális FNy szegmentumban.

- Konkordancia: amennyiben a fordítási környezet nem jelöl ki automatikusan egyes terminusokat, a fordító a fordítómemóriában megkeresheti a kérdéses kifejezések előfordulásait – mondatbeli környezetben. Így a fordítómemória segítségével mégis megláthatja, hogy ugyanazt a kifejezést a hálózatban dolgozó többi fordító hogyan fordította le.



3.7. ábra: A FNy terminológia kijelölése a fordítási környezetben

- Az aktuális részdokumentum fordításának befejezését jelezni kell a fordítástámogató rendszernek, amely ekkor automatikus teljesség-ellenőrzést végez. Ez azt jelenti, hogy a fordításból egyetlen mondat sem marad ki, így erre a kontrollszerkesztőnek egyáltalán nem kell figyelnie.

A terminológiai munkafolyamat ismertetésekor (az 5. fejezetben) részletesen kitérek a fordítási projektek különböző terminológiakezelési lehetőségeire. Az infrastruktúra ismertetéséhez azonban szükséges megemlíteni, hogy lehetnek olyan projektek, ahol a rendelkezésre álló idő nem teszi lehetővé a teljes terminológiai előkészítést. Ilyenkor a hálózatban elérhető terminológia különös jelentőséget kap. A konzisztencia biztosításához elengedhetetlen, hogy a projekt résztvevői terminológiai adatbázist építsenek. A fordítókat azonban a legtöbb projektben nem tekinthetjük a CNy terminológia autentikus forrásának, azon egyszerű oknál fogva, hogy többnyire nem szakértői a fordítandó szöveganyag tartalmának, és nem ismerik a témához tartozó terminológiát. A projektben terminológusként részt vevő szakértőnek azonban nincs lehetősége a teljes FNy szöveg elolvasására, különösen előzetesen nem. Ezért az új terminológia hozzáadása óhatatlanul a fordítókra hárul.

A MemoQ fordítási környezet fejlesztése során ezért kidolgoztam a moderált vagy felügyelt terminológiai adatbázis koncepcióját és a rá épülő rész-munkafolyamatot. A terminológus által felügyelt, fordítók között szétosztott terminológiakezelés nem újdonság, mint ahogy a terminológiai kérdéseket kezelő számí-

tógépes rendszer sem az. Előbbit rendszeresen alkalmaztuk 1998 és 2000 között az egyik EU-jogharmonizációs fordítási projekt terminológiai munkáiban, utóbbit szoftverhonosítással foglalkozó fordítóirodák rendszeresen alkalmazzák. Ezek hiányosságai alapján azonban meg lehetett tervezni egy olyan rendszert, amely minimalizálja a terminológia beviteléhez és ellenőrzéséhez szükséges munkát.

Ennek során minden fordító a hálózatban elérhető közös terminológiai adatbázist használja. A fordítási környezet lehetővé teszi, hogy mindenki, aki rendelkezik az ehhez szükséges hozzáférési joggal, egy lépésben, a fordításszerkesztő elhagyása nélkül vegyen fel új elemeket a terminológiai adatbázisba. Alapértelmezés szerint ezek az új tételek automatikusan hozzáférhetővé válnak a terminológiai adatbázis többi felhasználója, vagyis a többi fordító számára.

Mivel a fordító az alapfeltevés szerint nem szakértője a FNy szöveg témájának, az általa felvett terminológiai tételt nem lehet automatikusan helyesnek elfogadni. Az új terminológiára azonban gyorsan szükség van, ezért lehetővé kell tenni, hogy minimális munkával és minimális idő alatt ellenőrizni lehessen.

Az új terminológiai tétel ezért nem válik automatikusan elérhetővé a projekt többi szereplője számára, hanem felkerül egy speciális listára, amelyet a „terminológus" szerepben dolgozó felhasználó láthat. A terminológusnak be kell jelentkeznie a terminológiai adatbázis kiszolgálójára, ahonnan kap egy listát, amelyen a frissen felvett, ellenőrzésre váró tételek szerepelnek. Az egyes tételeket a terminológus egy lépésben elfogadhatja, elutasíthatja vagy javíthatja. A terminológus által véglegesített tételek azonnal és automatikusan láthatóvá válnak a fordítók számára.

U t ó f e l d o l g o z á s . A minőségbiztosítás technológiai segítése azt a célt is szolgálja, hogy az utóellenőrzés ideje, s így költsége csökkenjen. Erre három módszert ismertetek, amelyek közül csak az első közkeletű:

(1) Automatikus konzisztencia-ellenőrzés

(2) Szimultán lektorálás

(3) A lektorálás egyes műveleteinek automatizálása

Automatikus konzisztencia-ellenőrzés. A párhuzamos szöveg – a CNy szöveg és a mondatszinten hozzá igazított FNy fordítás – ismeretében az inkonzisztencia bizonyos felszíni jegyei felismerhetők. Ezek a felszíni jegyek – a teljesség igénye nélkül – a következők:

- Teljes szegmentum inkonzisztens fordítása: ugyanaz a szegmentum többször előfordul a FNy szöveganyagban, de a különböző helyeken eltérően fordították;
- Terminológia inkonzisztens fordítása: a terminológiai adatbázisban szereplő tétel FNy alakja szerepel a FNy szegmentumban, de a hozzá tartozó CNy szegmentumban nincs meg a terminológiai adatbázisban hozzárendelt CNy megfelelő.

- Tiltott terminológia alkalmazása: adott szó vagy frazéma FNy alakja mellett a terminológiai adatbázisban olyan CNy alak szerepel, amelyet a fordítás során nem szabad használni, a vizsgált CNy szegmentumban mégis szerepel.
- Nem fordítandóként megjelölt FNy karaktersorozat nem szerepel a vizsgált szegmentumhoz tartozó CNy szövegben.

A konzisztencia-ellenőrző (közkeletű nevükön minőségbiztosítási) eszközök emellett formai ellenőrzést is végeznek, így kiszűrik a számok, dátumok, pénzösszegek helytelen írását, a halmozott központozási jeleket vagy szóközöket is.

Szimultán lektorálás. Sok fordítási projekt olyan időpréssel néz szembe, amely mellett nincs lehetőség utólagos ellenőrzésre. Találkoztam olyan fordítási projekttel, ahol a lektornak (az iroda szóhasználata szerint „nyelvi vezető") 6 órát adtak 100 000 szónyi (kb. 600 000 leütésnyi) fordítás ellenőrzésére. Ez újabb bizonyíték arra, hogy a növekvő időprés a fordítások minőségének romlását hozza magával, függetlenül a minőségről alkotott fogalmunkkal – már ha elfogadjuk azt a hipotézist, hogy a lektorálás által javul a CNy szöveg minősége.

A technológia eszközeivel olyan megoldást kerestünk, amely a lektorálást akkor is lehetővé teszi, ha arra nincs idő. A konzisztencia biztosítása és a fordítás gyorsítása érdekében már korábban bevezettük a közös, a fordítók számára hálózaton keresztül elérhető fordítómemóriát. Nos, ha a fordítók a fordítási környezetet rendeltetésszerűen, állandó hálózati kapcsolat mellett használják, és minden szegmentumot rögtön lefordítása után jóváhagynak, a fordítás nemcsak a fordítóknál tárolt dokumentumban, hanem a közös kiszolgálón levő fordítómemóriában is létrejön, tehát a projekt többi szereplője – köztük a lektor – számára is azonnal hozzáférhető lesz.

A lektor a fordítás alatt álló FNy szöveget maga is megnyithatja a fordítási környezetben, és kérheti annak előfordítását a fordítási környezettől. Az előfordítás azt jelenti, hogy a fordítási környezet automatikusan, a felhasználó beavatkozása nélkül, egyenként megkeresi a közös fordítómemóriában a FNy szöveg szegmentumait, és az egyező FNy szegmentumok mellett tárolt CNy szegmentumokat beírja a munkaterületre, így kitöltve a CNy szöveget. Ezt követően a lektor elolvashatja és javíthatja a CNy szegmentumokat, és a javított CNy szöveget – a szegmentumok jóváhagyásával – visszavezetheti a közös fordítómemóriába – vagy egy másikba, amely a „lektorált" szöveget hivatott tárolni.

Ennek a munkamódszernek két hátránya van:

- A fordító kezéből automatikusan kiveszi az utólagos javítás lehetőségét (hiszen minden szegmentum „leadott" fordításnak számít a jóváhagyása pillanatában), pedig a fordítás során számtalanszor előfordul, hogy adott kisebb szövegelemet a későbbi szövegkörnyezet alapján lehet csak megérteni. Bár a visszalépést az időprés eleve lehetetlenné teheti, ez a probléma a megfelelő kommunikációs infrastruktúrával kiküszöbölhető: a fordító

jelezheti a lektornak, hogy javított a már átadott CNy szövegben, s ilyenkor a lektornak nincs más dolga, mint „lehívni" a közös fordítómemóriából a javítás által érintett szegmentumokat.

- Ha a fordítómemória csak a FNy szegmentumok egyezését figyeli – tehát az adatbázisból azt az első CNy szöveget tölti le, amelynél a FNy szegmentum megegyezik az aktuális FNy szegmentummal –, téves CNy szegmentumok is megjelenhetnek a fordításban. Ennek az az oka, hogy ugyanazon FNy szöveg több különböző környezetben is előfordulhat, és a különböző környezetek esetleg eltérő fordítást követelhetnek meg. Minél rövidebb az FNy szegmentum, annál nagyobb ennek a valószínűsége. Emiatt a szimultán lektoráláshoz olyan fordítómemóriára van szükség, amely nemcsak a FNy szegmentumok szövegét, hanem azok környezetét – folytonos szöveg esetén a megelőző és a következő FNy szegmentum szövegét – is feljegyzi, és a lekérdezés során ezt is figyelembe veszi.

A lektorálás egyes műveleteinek automatizálása. A fordításszöveg javítása közben is gyakran sablonszerű átváltási műveleteket kell végrehajtani, amelyek a forrásnyelvi formához és a szöveg témájához kapcsolódnak. Ilyenek egyes, a fordításszövegben megjelenő helyesírási hibák is. Ha ezt elfogadjuk, akkor kézenfekvő, hogy ezeket az átváltási műveleteket is párhuzamos korpuszok felhasználásával kell vizsgálni.

A korpusznyelvészet eszközeivel a jelenleginél sokkal többet megtudhatunk a fordítások minőségéről és értékeléséről is. Mona Baker (Baker 1996) ezt fordításszövegek korpuszán keresztül igyekszik vizsgálni, felismerve, hogy a fordításszövegek sajátos tulajdonságokat mutatnak az adott célnyelven írott eredeti szövegekhez képest.

Azonban fordításszöveg is többféle van. Amikor egy szöveget publikálásra készítünk elő, több végigolvasás során, iteratív módon javítjuk. Ugyanezt teszszük, talán még jobban definiált formában, amikor fordításszöveget készítünk elő megjelentetésre. A publikált fordításszöveg – különösen szakfordítások esetén – megfelelő szerkesztési protokoll esetén igen távol esik a fordító által átadott fordításszövegtől. Így kijelenthetjük, hogy nemcsak a FNy és a CNy szöveg között, hanem a kezdeti és a publikált CNy szöveg között is szükség van transzformációra. Ez utóbbit alkotják a lektorálás, olvasószerkesztés, korrektúra műveletei. A kommunikációs láncban ugyanakkor ez a másodlagos transzformáció integráns részét képezi a kibocsátó és a befogadó közötti, fordításnak nevezett külső transzformációnak.

A SZAK Kiadó műhelyében létrehoztunk egy speciális párhuzamos korpuszt, amely a fordítások első és publikált szövegeit tartalmazza egymás mellett. A fordítás és javításának együttes vizsgálatáról már volt szó a 3.1. fejezetben, az ekvivalencia problémájának tárgyalásánál.

Ezek a szövegváltozatok jól összehasonlíthatók egymással, nem úgy, mint a különböző nyelvű szövegeket tartalmazó párhuzamos korpuszokban levők.

Ennek oka elsősorban az, hogy mindkét komponens a célnyelven van leírva, másodsorban pedig az, hogy a lektorálás során általában nem kerül sor a szöveg teljes újrafogalmazására vagy újrafordítására. Ebben az esetben a szövegváltozatok egyszerűen összehasonlíthatók, még különösebb nyelvtechnológiai apparátusra sincs szükség. A korpusszal és a kutatás módszertanával részletesen a 4. fejezet foglalkozik.

4. A fordítástechnológia kapcsolata a korpusznyelvészettel és a nyelvtechnológiával

4.1. Általános megállapítások

Nyelvtechnológia és számítógépes nyelvészet

A fordítástechnológiával fennálló kapcsolat meghatározásakor a nyelvtechnológiát a számítógépes nyelvészet alkalmazásának tekintem. Értelmezésemben a számítógépes nyelvészet az írott szövegek és a beszéd számítógépes analízisével és szintézisével foglalkozik, különböző nyelvi szinteken és mélységben. Ezt a distinkciót sok számítógépes nyelvészettel foglalkozó személy és szervezet sem teszi meg.[21] Minden számítógépes nyelvészeti eljárás arra irányul, hogy a szövegről vagy annak elemeiről – szisztematikus módon – minél többet tudjunk meg, illetve a számítógépen tárolt strukturált információt minél jobban formált nyelvi produkcióval tudjuk megjeleníteni (vagyis szöveget szintetizálni belőle). A nyelvtechnológia pedig az így megszerzett ismereteket rendezi kézzelfogható eszközökbe, amelyek elsősorban számítógépes alkalmazások részeivé válnak.

Mivel a számítógépes nyelvészet kutatásait kezdetben a gépi fordítás létrehozása motiválta, a fordítástechnológiával szerves kapcsolatban is állhatna. A fordítástechnológiai eszközök ugyanakkor – sem a mikro-, sem a makrostratégiában – nemigen használják a nyelvtechnológia által létrehozott eszközöket, holott ezt számos szerző (pl. Hodász G. et al. 2004, Callison-Burch et al. 2005) javasolja. A fordítástechnológia számos olyan szövegkeresési és szövegmanipulációs eszközt alkalmaz, amelynek olyan értelemben nincs köze a nyelvtechnológiához, hogy nem veszi figyelembe, nem kísérli meg feltárni és manipulálni a szöveg nyelvi struktúráját.

Ha az ellenkező irányt tekintjük, a modern számítógépes nyelvészet és nyelvtechnológia számos területen profitál a fordítástechnológiából, mégpedig azért, mert a fordítás spekulatív, kvázi-kompetenciamodelljeivel szemben a megvalósult fordítás performanciaalapú modellt kínál a fordítás tanulmányozásához. Ez ugyanaz a megközelítés, mint a korpusznyelvészeté a nyelv modellezésével kapcsolatban. A fordítástechnológia felügyelete alatt végzett fordítás során jelentős, szegmentumszinten szinkronizált párhuzamos korpuszhoz jutunk, amelyek különböző szempontok szerint „bányászhatók". Már az ALPAC-jelentés (Pierce et al. 1966) és a XEROX (Kay 1980) is felvetette, hogy a spekulatív modellek helyett a megvalósult emberi fordítás újrahasznosításával kellene növelni a fordítás hatékonyságát. A kiterjedt, nagy tömegű párhuzamos korpu-

szok létrejötte mindenesetre elősegítette a statisztikai gépi fordítás létrejöttét (vö. pl. Callison-Burch et al. 2004).

A fordítástechnológia és a nyelvtechnológia viszonya azért ambivalens, mert a megfelelő minőségű nyelvtechnológiai eszközök kialakítása rendkívül költséges. Ez azért lényeges, mert a fordításhoz hasonlóan a fordítástechnológiai eszközök előállítása is gazdasági tevékenység, amelynek viszonylag alacsony a tőkebázisa, és viszonylag szűk maga a piac is. A Common Sense Advisory[22] szerint a fordítástámogató eszközök elterjedését már a jelenlegi formájukban is gátolja a viszonylag magas áruk. Vizsgálatuk szerint egy hagyományos fordítástámogató eszköz évente 700-2000 dollár közötti összegbe kerül egy fordítónak (új vásárlás, frissítés, terméktámogatás stb.). Ugyancsak a Common Sense Advisory (és mások, pl. a Gilbane Group) szerint a fordítástámogató eszközök piacának becsült nagysága évi 100 millió dollár.

A nyelvtechnológiai eszközök ráadásul mind nyelvspecifikusak: vagy nyelvfüggő szabályrendszerre, vagy jelentős méretű, nyelvre vagy nyelvpárra kidolgozott korpuszra van szükség. A fordítástechnológiai eszközök gyártói nem engedhetik meg maguknak, hogy termékeik csak egy-két nyelvre vagy nyelvpárra legyenek használhatók, így a nyelvspecifikus technológiát több tíz nyelvre vagy nyelvpárra is be kell szerezniük. A MorphoLogic tapasztalatai mutatták, hogy a szabályalapú gépi fordítás kifejlesztése egyetlen nyelvpárra is elérheti a 20 emberévet, ami lehetetlenné teszi a tisztán piaci alapú finanszírozást és megtérülést. Így, miközben a fordítástechnológiai eszközök gyakorlatilag bármilyen nyelvpárral vagy nyelvi együttessel használhatók, a nyelvtechnológiai eszközök a legtöbb nyelvhez vagy nyelvpárhoz nem is állnak rendelkezésre, a kifejlesztésükhöz szükséges tőke pedig nincs jelen a fordítástechnológiai piacon.

Korpusznyelvészet

A korpusznyelvészetet a számítógépes és leíró nyelvészet kísérleti eszközének tekintjük, mivel jó közelítő modellt szolgáltat a nyelv viselkedéséhez. A fordítástechnológia e tekintetben közvetlen adatforrás a korpusznyelvészet számára, hiszem – mint már említettem – a technológia felügyelete mellett végzett fordítás automatikusan párhuzamos korpuszt eredményez. A fordítási környezetek emellett rendszerint tartalmaznak két olyan eszközt, amelyet magam a korpusznyelvészetből ismertem meg: ez a szövegszinkronizáló *(aligner)* és a konkordanciaprogram *(concordancer)*; az egyik a párhuzamos korpuszok létrehozását, a másik pedig a korpuszon végzett kutatást segíti.

Bár a nyelvtechnológia és különösen a gépi fordítás kutatása igen intenzíven használja a párhuzamos korpuszokat, a fordítástechnológia relatíve keveset profitál belőlük – a triviális felhasználási módjuk mellett igazából semmit. A triviális felhasználási mód a fordítómemória, amely teljes szegmentumok megkeresésére képes a teljes vagy a részleges egyezés alapján. Ennek elterjedt algoritmusa a fuzzy keresés *(fuzzy search)*, amely a korpuszstatisztika bevett módszereihez hasonlóan betűkettesek és -hármasok (digráfok és trigráfok), illetve

szókettesek és -hármasok együttes előfordulásának számolásával határozza meg két FNy szegmentum hasonlóságának mértékét. A fuzzy logikához csak annyiban van köze, hogy az adatbázisban talált és a FNy szövegben levő FNy szegmentumok hasonlóságának mértékét egy 0 és 1 közötti igazságértékkel jellemzi, amelyet a legtöbb eszköz %-ban fejez ki.

A párhuzamos korpuszok kihasználása a korpusznyelvészet speciális ága – lehetne, amelyből a fordítástechnológia, azon belül pedig a fordítómemória-használat közvetlen hasznot húzhat. Szükséges a fordítómemóriák kihasználásának mértéke, illetve olyan kutatás folytatása, amellyel ez a kihasználtság növelhető.

Párhuzamos korpuszok és szövegszinkronizálás a fordítástechnológiában

A párhuzamos korpusz olyan értelemben másodlagos fordítási erőforrásnak számít, hogy a fordításhoz közvetlenül sem a gépi fordítástámogató, sem a gépifordító-rendszerek nem tudják felhasználni.

Ugyanakkor elsődleges fordítási erőforrás annyiban, hogy – mivel emberi fordításokat tartalmaz – mindenképpen autentikus. Erre alapoz a statisztikai és a determinisztikus példaalapú gépi fordítás is (pl. Callison-Burch et al. 2004, Brown et al. 1994, Matusov et al. 2005).

A párhuzamos korpusz definíciójába beleértem a szinkronizáltságot is. Mivel mind a gépi fordítástámogatás, mind a gépi fordítás a forrásszöveget szegmentumokra bontva dekomponálja a fordítási feladatot, a párhuzamos korpusz is csak akkor használható fel, ha a benne levő forrásszövegek egyes szegmentumai meg vannak feleltetve a célszöveg megfelelő szegmentumainak.

E megfeleltetés létrehozásában – a szinkronizálásban – láthatjuk a párhuzamos korpuszok létrehozásának munkaigényét. Ezért régóta fennálló kutatási probléma a szövegszinkronizálás automatizálása. Ezért a hazai és nemzetközi szakirodalomban számos megközelítés és algoritmus olvasható (Gale-Church 1994, Pohl 2004).

Ennek a dolgozatnak nem tárgya a szövegszinkronizálási algoritmusok ismertetése, viszont a szinkronizálási munkát meg kell vizsgálni a fordítástechnológia szempontjából.

A szinkronizálásnak számos szintje van, ezek közül azonban nem mind hasznos a fordítástechnológiában. A gépi fordítástámogatás a szinkronizálás következő szintjeit tudja felhasználni:

- Mondatszinkronizálás: ekkor a fordítómemóriák fordítási egységeiben szereplő szegmentumokat feleltetjük meg egymásnak, tehát a szinkronizálással tulajdonképpen fordítómemóriát hozunk létre.
- Terminológiakeresés: a párhuzamos korpuszok segítségével, statisztikai módszerek felhasználásával egyes terminusok célnyelvi megfelelőit keressük. A keresés alapja az, hogy a forrásnyelvi kifejezés forrásszövegbeli eloszlását alapul véve olyan szavakat, kollokációkat keresünk a célszövegben, amelyek

gyakorisága és eloszlása megfelel a forrásnyelvi kifejezésnek. (Blank 2000, Choueka et al. 2000, Callison-Burch et al. 2005)

- Szegmentum alatti részek (például főnévi csoportok) szinkronizálása: nyelvi támogatású fordítómemóriák feltöltése esetén erre is szükség van, mivel a forrás- és célszegmentumok dekompozíciója után a kisebb építőelemeket meg kell feleltetni egymásnak. (Pohl 2006)

A szinkronizálási algoritmusok valamennyien kötegelt üzemmódú – nem interaktív – automatikus végrehajtásra készültek, ám nem teljesen megbízhatók. A fordítástechnológiának – a makrostratégia szintjén – ezért szüksége van olyan interaktív alkalmazásokra, amelyekkel az ember az automatikus szinkronizálás eredményét javítani tudja. Alább ennek – és a fenti három szinkronizálási szintnek – a bővebb ismertetése következik.

A fordítómemória maga is párhuzamos korpusz, hiszen forrásszövegek szegmentumait tartalmazza, célszövegek szegmentumainak megfeleltetve. A korpuszjelleg ott sérül, hogy a fordítómemóriában tárolt szövegek nem intaktak, vagyis az egyes forrásdokumentumok nem mindig rekonstruálhatók az adatbázisból: oda szegmentumokra bontva kerülnek, és a szövegszervezési elvektől eltérő alapokon vannak rendezve.

A fordítómemória kétféleképpen tölthető fel:

- interaktív (felügyelt) fordítási folyamat során, illetve
- párhuzamos szövegek szinkronizálása útján.

E szempontból az egyetlen értelmes szinkronizálási szint a szegmentumoké. A szinkronizálás során ráadásul ugyanazt a szegmentálási algoritmust kell alkalmazni, mint a fordítómemória működése közben, különben a szinkronizálással bevitt szegmentumok nem lesznek minden esetben megkereshetők a fordítás közben.

A szegmentálás azért okoz problémát, mert a fordítástámogató eszközök fejlesztői a szöveg fordítási egységének kénytelenek a mondatot választani, amelynek gépi elhatárolása viszonylag egyszerű, nem igényel sok nyelvfüggő adatot. A mondatszegmentálás ugyanakkor nem is egyértelmű és nem is tökéletes. A mondathatárokat különböző programok másképp értelmezik, sőt, a szegmentálási szabályok egy programon belül is megváltoztathatók.[23] Ezért a szinkronizálás során figyelemmel kell lenni a szegmentumhatárokra.

Sok fordítási környezetet, köztünk az általunk kifejlesztett MemoQ rendszer is tartalmaz szinkronizáló modult, amelyek működése három fázisból áll:

- a forrás- és a célszöveg szegmentálása,
- a szegmentumok automatikus szinkronizálása,
- a szinkronizálás manuális, interaktív és sok esetben iteratív javítása.

Az iteratív javítás abban az esetben használható, ha az automatikus szinkronizálási algoritmus az ember által megadott biztos szinkronizálási pontokat, az úgynevezett h o r g o n y o k a t *(anchors* – lásd: Pohl 2004), figyelembe tudja

venni a szinkronizálás végrehajtásánál. Ekkor a szinkronizálási munka harmadik fázisa úgy zajlik, hogy az ember megjelöl egy vagy két biztos szinkronizálási pontot, majd újra futtatja az automatikus szinkronizálást.

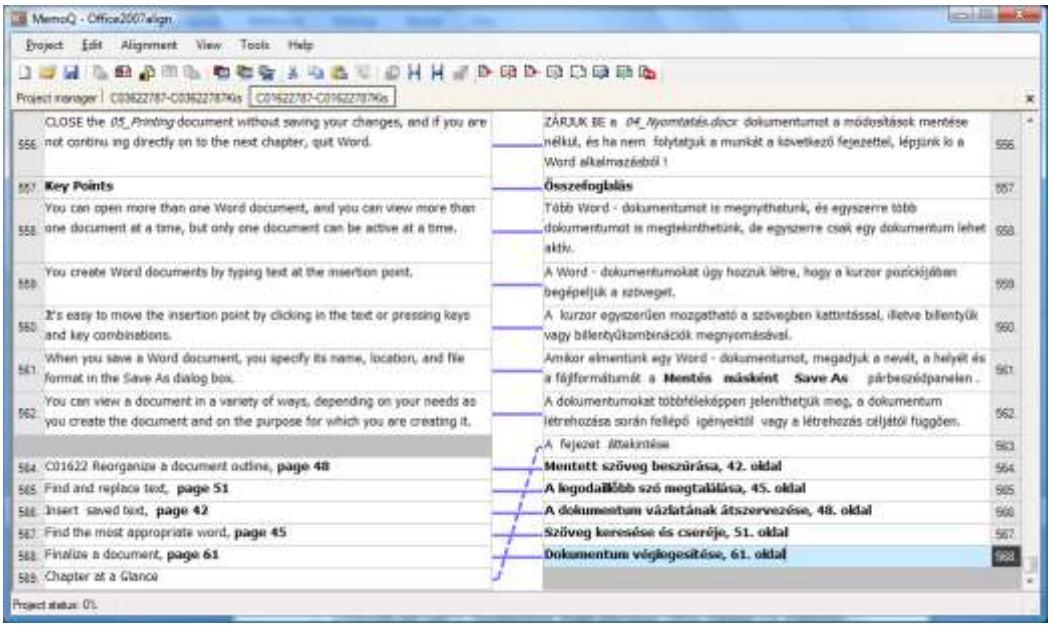
Ennek kapcsán össze kell foglalni az automatikus szegmentum- (mondat-) szinkronizáló algoritmusok működését. Ezek, ha már megtörtént a forrás- és célszöveg szegmentálása, a következőképp feletetik meg egymásnak az egyes szegmentumpárokat:

- Horgonyok alapján (pl. Kay-Röscheisen 1994, Pohl 2004): ha ugyanaz a lexikális vagy épp extralingvális információ a forrás- és a céloldalon is megvan egy szegmentumban, akkor azokat egymáshoz tartozónak tekintik. Ilyen horgonyok a le nem fordított nevek, a számok, illetve a (terminológiai vagy általános) szótárban egymásnak megfelelő szavak. (Ez utóbbi felhasználása esetén az algoritmusnak természetesen szótárt is kell használnia.)
- A szegmentumok hossza alapján (Gale-Church 1994): ekkor azt feltételezik, hogy az egymáshoz rendelt forrás- és célszegmentum hosszának összemérhetőnek kell lennie (egy toleranciaküszöbön belül). Ennek alapján lehet észrevenni, ha a fordítás során összevontak több forrásszegmentumot vagy szétválasztottak egyet. Az ilyen algoritmusok képesek szegmentumok automatikus összevonására.

Az irodalomból ismert valamennyi szinkronizálási algoritmus feltételezi a forrásszegmentumok sorrendtartó lefordítását, ami viszont valós fordítások esetén nem mindig érvényes – igaz, nem is nagyon gyakori. Sok esetben a szövegdobozokat tartalmazó dokumentumok eltérő tördelése okozza a sorrendcserét, időnként azonban a FNy szövegben több bekezdésre tagolnak hosszú mondatokat tagoltak, s ekkor a szegmentáló algoritmus mondatnál jóval kisebb egységeket tekint szegmentumnak. Azonban ezt – legalábbis karaktersorozat-alapú fordítómemória használata esetén – nem tanácsos úgy korrigálni, hogy több bekezdéses mondatokat is megengedünk, mert ennek nagyon hosszú szegmentumok lesznek az eredményei, ami pedig jelentősen csökkenti a fordítómemória hatékonyságát.

A fordítástechnológia kettős célja a fordítási munka hatékonyságának növelése és a fordítások minőségének javítása. Ha e szempontból vizsgáljuk a fordítómemória építésére irányuló szinkronizálást, akkor annak hatékonyságával is foglalkoznunk kell. A fordítómemória által eredményezett hatékonyságjavulás ugyanis elveszhet, ha a szinkronizálás túl sok emberi munkát emészt fel.

Emiatt nemcsak a szinkronizálási algoritmusok kutatását kell a fordítástechnológia kutatási területei közé sorolni, hanem a szinkronizálási munkafolyamat értékelését, javítását is. Ennek már nemcsak az automatikus szinkronizálást kell értékelnie, hanem a hozzá kapcsolódó szegmentálási és utójavítási infrastruktúrát is – ez utóbbinak egyelőre nincs irodalma [még az oly gyakorlatias Austermühl (2001) sem foglalkozik vele!], miközben az automatikus szinkronizálásnak elfogadott és bevezetett elméletei vannak.



4.1. ábra: Felhasználói felület a szinkronizálás utószerkesztésére a MemoQ programban

Az összetett szinkronizálási eljárások hatékonyságának mérése még kidolgozásra vár, de már ezen a ponton is nyilvánvaló, hogy a következő paramétereket mindenképpen figyelni kell:

- az automatikus szinkronizálás után szükséges korrekciós műveletek (összerendelés törlése, új összerendelés létrehozása) számát, illetve arányát a szöveg terjedelméhez viszonyítva: ez a szinkronizáló algoritmus minőségét mutatja;
- az utókorrekció során szükségessé váló szegmentumkorrekciós műveletek (összevonás, szétválasztás) gyakoriságát, illetve arányát a szöveg terjedelméhez viszonyítva: ez a szegmentáló algoritmust minősíti, bár a hossz alapú szinkronizálás sok szegmentálási hibát kijavít;
- a korrekciós műveletek végrehajtásához szükséges időt: ez a felhasználói felület minőségét értékeli.

A fenti paraméterek mérése szoftverergonómiai módszerekkel lehetséges, olyan programok segítségével, amelyek lehetővé teszik a felhasználó tevékenységének feljegyzését, és mérik a műveletek idejét is. Ilyenek a billentyűzésfigyelő *(keylogger)* programok, amelyeket sokan kémprogramként használnak, ezért ma már elsősorban a biztonsági szakemberek foglalkoznak velük, nem pedig a szoftverergonómia kutatói.[24]

A folyamatnak a felhasználói felület mindenképp része, ezért a fenti szempontok szerint a „csupasz" szinkronizáló algoritmusok nem értékelhetők.

A fordítómemóriák szinkronizálással való felépítését – annak nagy munkaigénye, illetve a rendelkezésre álló algoritmusok megbízhatatlansága miatt – sokan gazdaságtalannak tekintik, ezért vannak olyan fordítástámogató rendszerek, amelyek fordítómemóriát nem, csak párhuzamos korpuszt kezelnek. Ilyen rendszer például a MultiTrans.

Ezek esetében a szinkronizálást a keresés közben, sokszor manuálisan hajtjuk végre. A keresés menete a következő lehet:

(1) A fordító kijelöli a forrásszövegben az aktuális forrásszegmentumot vagy annak egy részét, majd indítja a keresést a fordítástámogató rendszerben.

(2) A fordítástámogató rendszer megjeleníti a kijelölt forrásszöveg (adott esetben közelítő) előfordulásait, és a célnyelvi szövegből a hasonló pozíción szereplő részt. A megfelelő fordítást a fordító keresi meg.

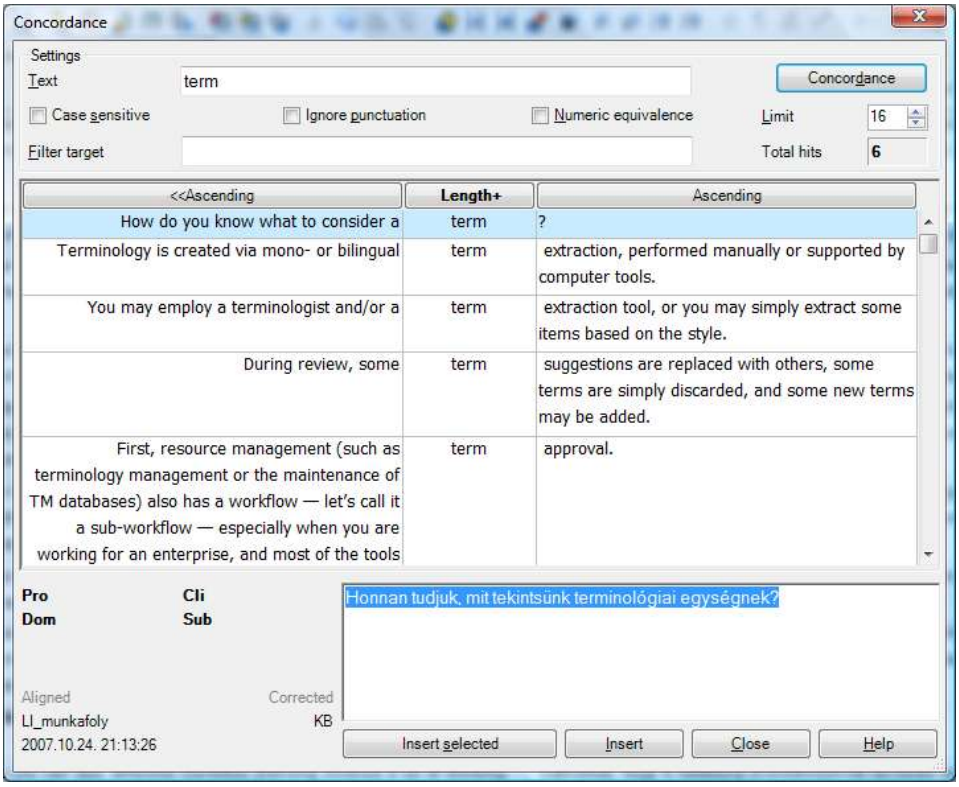
Fontos, hogy ebben az esetben a párhuzamos korpusz nincs szinkronizálva.

Megfigyelésem szerint az előkészített fordítómemóriák használata nagyobb hatékonyságot eredményez, de ez a módszer jól alkalmazható azokban az esetekben, amikor nincs mód vagy idő a párhuzamos szöveg előzetes szinkronizálására.

A fordítómemóriák konkordanciafunkciója

A fentihez hasonló szerepet a fordítómemória is betölthet (általában be is tölt). Ha az aktuális forrásszegmentumhoz nincs találat, de feltételezzük, hogy a forrásszegmentum egyes részei máshol előfordulhat, a legtöbb fordítómemória-rendszertől kérhetjük a kijelölt rész konkordanciáját.

Ekkor információt kapunk arról, hogy az – esetleg fordítási nehézséget jelentő – aktuális kijelölt szöveg még milyen környezetben fordult elő. Ez máris segítheti a megértését; azonban a fordítómemória a talált szegmentumok fordítását is rögtön felajánlja, amelyből fordítási javaslatot is kiemelhetünk.



4.2. ábra: Példa konkordanciára a MemoQ programban

A bemutatott kutatások

A következőkben három kutatást mutatok be, amelyek mindegyike a párhuzamos korpuszok kihasználását célozza a fordítástechnológiában:

- A SZAK javításkorpusszal lehetővé válik a szakfordítások lektorálási folyamatának modellezése. A 4.2. rész bemutatja a korpusz kialakítását és felépítését, és kutatási tervvel is szolgál a korpuszból nyert adatok feldolgozására.
- Az intelligens fordítómemória a nyelvi struktúrát is figyelembe veszi a párhuzamos korpuszban való keresés során. A 4.3. rész az ezzel kapcsolatos kutatást mutatja be (vö. Kis et al. 2004).
- A leghosszabb részszöveg konkordanciája és a fordítás automatikus összeállítása olyan eljárások, amelyek alkalmasak lehetnek több fordítómemória- és terminológiai tétel kombinálására oly módon, hogy olyan FNy szegmentumokhoz is használható fordítási javaslatot hozzanak létre, amelyekhez a „hagyományos" fuzzy keresés nem ad találatot. Ezek fejlesztése az értekezés írása idején éppen csak elkezdődött.

4.2. A SZAK javításkorpusz

A 3. fejezetben több helyen is esik szó arról, hogy a fordítás javításának legjobb modellje a megvalósult javítások elemzése. A javítások elemzéséhez a fordító által előállított CNy szöveg és a publikált CNy szöveg közötti különbséget kell vizsgálni. Ehhez olyan párhuzamos korpuszt kell létrehozni, amelyben a FNy és a publikált CNy szövegek helyett a CNy szöveg első változata és a publikált CNy szöveg található.

A SZAK Kiadó műhelyében létrejött egy ilyen korpusz, amely informatikai szakkönyvek és weblapok szövegét tartalmazza. A korpusz valójában három komponensből áll: az FNy szövegből, az első CNy szövegből és a publikált CNy szövegből.

A párhuzamos korpusz szegmentumszintű (mondatszintű) szinkronizálással állt elő; a szinkronizálást a MemoQ fordítási környezetbe épített szinkronizáló programmal végeztük. [Magyar nyelven a szövegszinkronizálásról lásd bővebben: Pohl (2004)]. A szinkronizált szöveg fordítómemóriába, onnan pedig a szabványos TMX formátumba került (Melby 2000).[25]

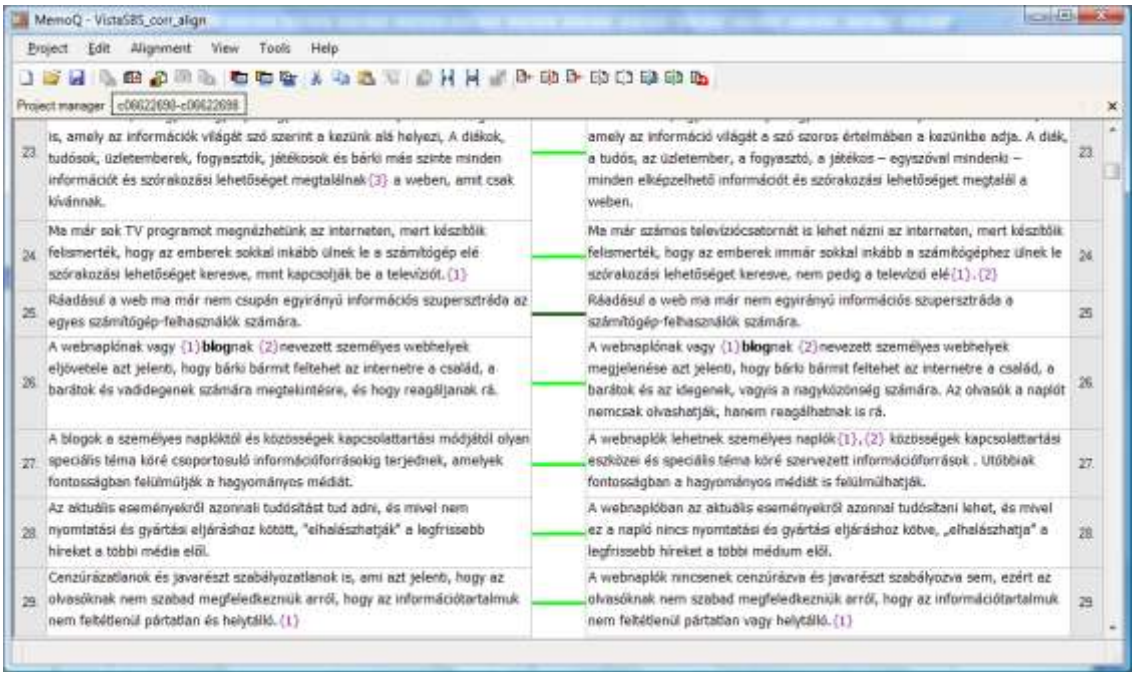
A korpusz két párhuzamos részkorpuszból áll össze: az egyik a FNy és a publikált CNy szöveg, a másik az első CNy és a publikált CNy szöveg szinkronizálásával keletkezett.

A korpusz mennyiségi és formai jellemzői

A SZAK javításkorpusz az értekezés írása idején 13 összetett, informatikai tárgyú írásmű szövegét tartalmazta. A terjedelmi adatok a következők:

	Szegmentumok száma	*Szövegszavak száma*
Forrásnyelven	139 508	1 473 519
Célnyelven	135 760	1 291 391

A szegmentálás, a tokenizálás és a szavak megszámlálása a MemoQ fordítási környezetben történt. A javítás előtti és a javított szöveg szinkronizálása ugyancsak a MemoQ fordítási környezet szinkronizáló rendszerében történt. A MemoQ rendszerre nemcsak azért esett a választás, mert magunk fejlesztettük, hanem azért is, mert a szinkronizáló rendszerhez kapcsolódó grafikus kezelőfelületen igen gyorsan észrevehetők és kijavíthatók az esetleges szinkronizálási hibák.



4.3. ábra: A javításkorpusz szinkronizálása a MemoQ rendszerben

A javítások elemzése azonban nem lehetséges a MemoQ-környezeten belül. A szinkronizálás eredményeképp kapott fordítómemóriát ezért a TMX formátumba mentettük, amely nem más, mint az XML nyelv egy részhalmaza fordítómemóriák ábrázolására.

<tu changedate="20080131T201142Z" changeid="KB">
<prop type="client"> </prop>
<prop type="project"> </prop>
<prop type="domain"> </prop>
<prop type="subject"> </prop>
<prop type="corrected">no</prop>
<prop type="aligned">yes</prop>
<tuv xml:lang="cs">
<prop type="x-context-pre"><seg>Ráadásul a web ma már nem csupán egyirányú információs szupersztráda az egyes számítógép-felhasználók számára.</seg></prop>
<prop type="x-context-post"><seg>A blogok a személyes naplóktól és közösségek kapcsolattartási módjától olyan speciális téma köré csoportosuló információforrásokig terjednek, amelyek fontosságban felülmúlják a hagyományos médiát.</seg></prop>
<seg>A webnaplónak vagy <bpt i='1' type='bold'>{}</bpt><ph type='fmt'>{}</ph>blog<ept i='1'>{}</ept>nak <ph type='fmt'>{}</ph>nevezett személyes webhelyek eljövetele azt jelenti, hogy bárki bármit feltehet az internetre a család, a barátok és vadidegenek számára megtekintésre, és hogy reagáljanak rá.</seg>
</tuv>
<tuv xml:lang="hu">
<seg>A webnaplónak vagy <bpt i='1' type='bold'>{}</bpt><ph type='fmt'>{}</ph>blog<ept i='1'>{}</ept>nak <ph type='fmt'>{}</ph>nevezett személyes webhelyek megjelenése azt jelenti, hogy bárki bármit feltehet az internetre a család, a barátok és az idegenek, vagyis a nagyközönség számára. Az olvasók a naplót nemcsak olvashatják, hanem reagálhatnak is rá.</seg>
</tuv>
</tu>

4.4. ábra: Egy fordítási egység (szegmentumpár) reprezentációja a TMX formátumban

A javítási folyamat rekonstrukciója

A korpusz a javítási folyamatot fekete dobozként jeleníti meg, mert csak a javítás kiindulópontját és végső eredményét látjuk benne. Az eredmény lehet több javítási fázis eredménye is, arra tehát nincs lehetőségünk, hogy az egyes közreműködők által végzett javítási műveletet rekonstruáljuk.

Ha azonban feltételezhetjük, hogy az egyik szöveg a másikból javítással, vagyis nem elvetéssel és teljes újraírással keletkezett, akkor segítségül hívhatjuk Levenstejn (1965) eljárását, amely eredeti formájában két rövid szövegrész egymástól való szerkesztési távolságát határozza meg, vagyis azon elemi műveletek számát, amellyel az egyik szöveg átalakítható a másikba.

A javítandó és a javított CNy szövegek közötti különbséget hasonló eljárással határozzuk meg. Levenstejn algoritmusához képest a követelmény három ponton változik:

(1) A karakterek helyett a szavakat kell egységnek tekinteni, de a részleges szóegyezést (vagyis a szójavítást) is figyelembe kell venni. Erre azért van szükség, mert az ortográfiai szó szintjén áttekinthetőbben ábrázolhatók a javítások, arról nem is beszélve, hogy a szokásos szegmentumhossz (20-100 karakter) mellett az algoritmus működése gyorsabb lesz.

(2) Nem a javítási műveletek száma, hanem tartalmuk az érdekes. A javítások mennyiségét ugyanakkor fel lehet használni a fordítások értékeléséhez. Az itt javasolt algoritmus a Levenstejn-eljárással szemben nem állítja elő automatikusan a szerkesztési távolságot, de a felismert javítási műveletek megszámolásával az is előállítható.

Érdemes megjegyezni, hogy az eredeti Levenstejn-algoritmus szövegek hasonlósági keresésére is alkalmas, a fordítómemóriák fuzzy keresési algoritmusaiban azonban nem használják, mert teljesítménye alacsony ahhoz, hogy egy 100 000 szegmentumnyi adatbázisban 1 ms körüli idő alatt megtalálja a keresett szöveghez leginkább hasonló szegmentumot.

(3) Nem elegendő a három szerkesztési alapművelet. Levenstejn eljárása három alapműveletet azonosít, a beszúrást, a törlést és az átírást. A fordítás javítása során azonban gyakori a szórend átrendezése, vagyis egyes szavak áthelyezése a mondaton belül más pozícióra. A módosított algoritmusnak ezért képesnek kell lennie az áthelyezés érzékelésére is.

Lássuk az algoritmust egy egyszerű (mesterséges) példán! (A korpusz valódi mondatai túl hosszúak ahhoz, hogy a javításukat áttekinthető módon be lehessen mutatni.)

- A javítás előtti CNy szöveg: *'*Az ebet viszem ma este sétálni.'*
- A javított CNy szöveg: *'Este leviszem a kutyát sétálni.'*

0. lépés: Tokenizálás. A szöveget tokenekre (szavakra) kell bontani. A központozási jelek és a mondatzáró írásjel külön tokennek számít. Legyen a javítás előtti CNy szöveg tokenjeinek száma n, a javított CNy szöveg tokenjeinek száma m.

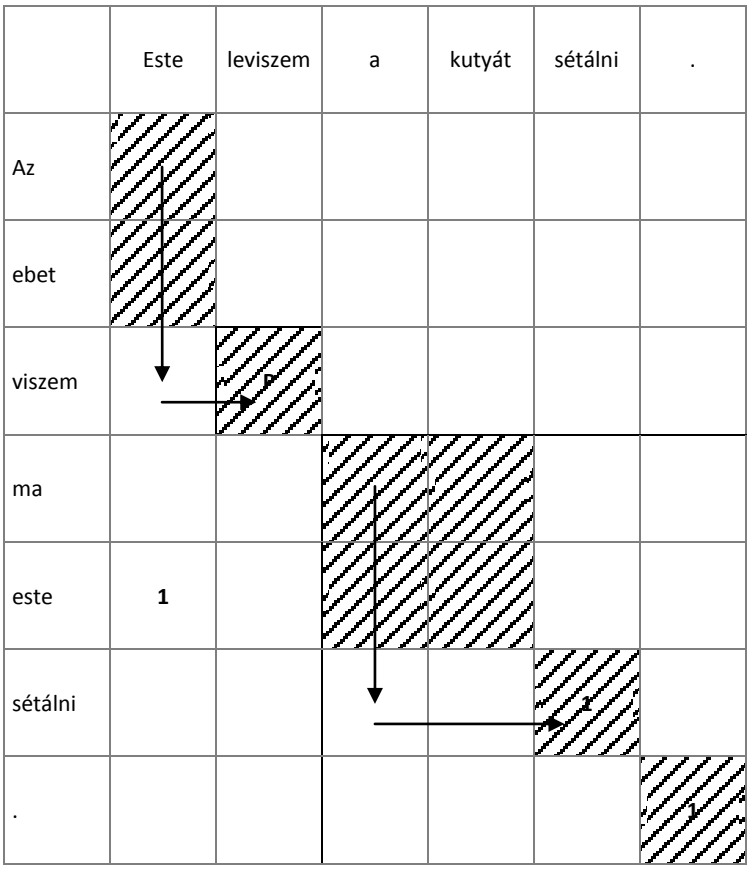
1. lépés: Ritka mátrix felírása. A következő lépésben fel kell venni egy $m \times n$ mátrixot, amelynek sorai a javítás előtti, oszlopai pedig a javított szöveg tokenjeivel vannak megcímkézve. A javítás előtti és a javított szöveg szavait össze kell hasonlítani, és ahol egyezést találunk, azt a helyet jelezni kell a mátrixban. A teljes szóegyezést az **1** számjegy, a részleges szóegyezést a **P** betű jelöli:

	Este	leviszem	a	kutyát	sétálni	.
Az						
ebet						
viszem		P				
ma						
este	1					
sétálni					1	
.						1

4.5. ábra: A módosított Levenstejn-algoritmus összehasonlítási mátrixa

2. lépés: A mátrix bejárása. Járjuk be a mátrixot a következő szabályok szerint:

a) Fölülről lefelé keressük meg az első sort, amelyben van egyezés, majd
b) ebben a sorban balról jobbra haladva keressük meg az első oszlopot, vagyis azt a cellát, amelyben legalább részleges egyezést találunk!
c) A lefelé mozgás során jegyezzük meg, mely sorokat, a jobbra mozgás során pedig azt, hogy mely oszlopokat léptünk át! Jelöljük meg az ezek által meghatározott területet!
d) Ismételjük meg a műveletet, de már csak a mátrixnak azon a részén, amelynek bal felső cellája a b) lépésben megtalált cella jobb alsó átlós szomszédja!



4.6. ábra: A módosított Levenstejn-algoritmus mátrixának bejárása

3. lépés: A javítási műveletek kiolvasása. Olvassuk ki a szószintű manipulációs műveleteket a mátrixból! Először csak a törléseket és a beszúrásokat kapjuk meg. A szabály: az egyezés nélkül elhagyott sorok által jelzett szavakat (tokeneket) törölték, az egyezés nélkül elhagyott oszlopokkal jelzett tokeneket pedig beszúrták. Ennek alapján a fenti mátrixból a következő lépéssor olvasható ki:

Törlés:	[Az] [ebet]
Beszúrás:	[Este]
Egyezés:	[(le)viszem]
Törlés:	[ma] [este]
Beszúrás:	[a] [kutyát]
Egyezés:	[sétálni] [.]

A bejáráskor szabály, hogy az egyezések után mindig átlósan kell jobbra lefelé haladni, amíg a mátrix lefelé vagy jobbra véget nem ér. Ha azonban a mondatzáró írásjel megegyezik, akkor a bejárás mindig a jobb alsó sarokban ér véget.

4. lépés: A szavak javításainak feltárása. A részlegesen egyező szavak javítási műveleteit hasonlóképpen, már a szavak karaktereire lebontott eljárással kell feltárni.

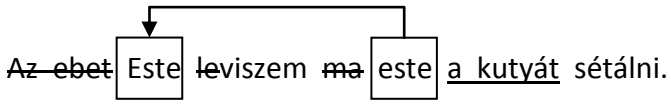
5. lépés: Az áthelyezés és az átírás feltárása. A fenti eljárás csak a beszúrást és a törlést ismeri fel, azonban nekünk két további művelet felismerésére is szükségünk van: ez az áthelyezés és az átírás. Az átírást egyszerűbb felismerni: minden olyan egymást követő törlés és beszúrás átírásként kezelendő, ahol a törlés és a beszúrás között nem volt egyezés.

Ennek felismerése azonban túlságosan mohó, és elfedi az áthelyezéseket. Az áthelyezések felismerésére szolgál a 4.6. ábrán besatírozott rész, amely a bejárás során érintett cellákat jelöli: ez a lineáris bejárás által lefedett terület. Ha van olyan egyezés, amely kívül esik a besatírozott területen, az áthelyezést jelez. A példában ilyen az *'este'*, amely a mondat végéről az elejére került.

Az átírások és az áthelyezések felismerése után a szerkesztési műveletek felsorolása a következőképpen fest:

Törlés:	[Az] [ebet]
Áthelyezés (cél):	[Este] ⑤→①
Egyezés:	[(le)viszem]
Törlés:	[ma]
Áthelyezés (forrás):	[Este] ⑤→①
Beszúrás:	[a] [kutyát]
Egyezés:	[sétálni] [.]

Ez grafikusan a következőképp ábrázolható:



A különbségek feldolgozása, a korpusz felhasználása kutatáshoz

A korpusz két, egy elméleti és egy gyakorlati célra használható fel: a javítások szabályosságainak vizsgálatán keresztül tanulmányozhatók a fordításjavítás folyamatai, áttételesen pedig az ekvivalencia fogalma. Másfelől pedig cél olyan számítógépes segédeszköz kidolgozása, amely képes a korábbi lektori javítások által érintett problémák észlelésére és – a javítások visszajátszásával – automatikus javítására. Ez a kutatás azonban csak most kezdődik, így a későbbi kutatókra vár a következő kérdések megválaszolása:

(1) Milyen nyelvi szinteket érintenek a javítások?

(2) Az ekvivalenciaszintek elmélete alapján változik-e az ekvivalencia szintje a javítások során, és ha igen, hogyan?

(3) A javítások elemzése alapján lehetséges-e olyan eszköz kifejlesztése, amely lehetővé teszi a javítási szituációk automatikus észlelését és a CNy bemenet automatikus javítását?

Az automatikus javítási eszköznek leginkább a lektormemória nevet adhatjuk. A lektormemória fejlesztéséhez elegendő az átírási és törlési műveletek vizsgálata; az áthelyezési műveletek felismerését – bár a konkrét javítások alapján észlelhetők – megnehezíti, hogy az áthelyezési forrás- és célpozíciók nem állapíthatók meg egységesen. A szöveg-szöveg cserék azonban alkalmasak olyan szabályosságok megállapítására, amelyek nyomán a korpuszból kinyert felszíni szövegek egyes részei szimbólumokkal helyettesíthetők, vagyis a szöveg javítására speciális reguláris kifejezések dolgozhatók ki. (A reguláris kifejezésekről lásd pl. Friedl 2003)

4.3. A fordítómemóriák értékelése és kihasználásuk javítása

A fordítómemória definíciója és motivációja

A fordítások újrahasznosítása. A fordítómemória a legelterjedtebben használt fordítási erőforrás, vagyis olyan számítógépes eszköz, amely a forrásszöveg egészére vagy egy részére valamilyen szisztematikus eljárással fordítást javasol. A fordítómemória tulajdonképpen párhuzamos korpusz, amely úgynevezett fordítási egységekben egymás mellé rendelve tartalmazza egy vagy több dokumentum FNy és CNy szegmentumait.

Az emberi fordítási munka során nagy tömegű fordítás keletkezik. Emellett felismerhető, hogy egyes szövegtípusokban gyakori a belső és a külső ismétlődés: egyes szövegrészek sablonszerűen újra és újra megjelennek a forrásszövegben.

Belső ismétlődésnek az egyes részeknek az integráns szövegen belüli ismétlődését nevezzük, a külső ismétlődés pedig a szövegközi ismétlődés, amely egy adott műfajhoz, tárgykörbe vagy adott forrásból (műhelyből) származó szövegek halmazán belül érvényes.

A különböző tárgykörök szövegei, illetve a különböző szövegtípusok ismétlődéskarakterisztikája szélsőségesen eltérhet. A közvélekedés szerint a hosszabb részek sablonszerű ismétlődése leginkább jogi szövegekre jellemző, de tipikusnak mondjuk a műszaki leírások szövegeit is. Közepes terjedelmű (5000-7000 szegmentumból álló, 60-70 ezer szövegszónyi) informatikai szakkönyvekben a *belső* ismétlődés 5-10%.

Ha egyes, nagy korpuszok által képviselt műfajokra és tárgykörökre jellemző a belső és külső ismétlődés, akkor az ilyen szövegek fordítása jelentősen felgyorsítható az ismétlődések kihasználásával. Ezek jellemzően szakmai szöve-

gek – s minél jellemzőbb az ismétlődés, annál inkább formalizáltak –, amelyek stilisztikai értéke és igénye eleve kisebb, mint a narratívumoké. A gyorsítás abban áll, hogy a fordítónak nem kell újra lefordítania olyan szövegrészt, amely ugyanazon szövegben, illetve a fordítói praxisban korábban már előfordult.

Ez a gyorsítás számítógép nélkül lehetetlen, bár vannak olyan fordítók, akik jól emlékeznek egyes, ismétlődő szövegrészekre, s azokat hatékonyan meg tudják keresni. Azonban a legtöbb ismétlődés esetén a manuális visszakeresés idő- és munkaigénye általában összemérhető az újrafordításával (minél rövidebb szövegrészről van szó, annál inkább).

A fordítómemória mint számítógépes eszköz rendeltetése a fordító vagy fordítócsoport korábbi fordításai között is előforduló szövegrészek gyors felismerése és megkeresése, illetve a korábban adott fordítás visszaadása valamely aktuális forrásszöveg fordítása közben. Ehhez a fordítómemóriának olyan adatbázist kell fenntartania, amelyben – megfelelő felbontással – tárolva vannak a korábbi forrásszövegek és fordításaik.

A fordítómemóriának ezért a fordítási folyamat során végig működnie kell, mintegy felügyelve a fordító munkáját (ez a felügyelt fordítás). Ezzel lehetséges, hogy a jól szinkronizált fordítási egységek már a fordítás közben kialakuljanak, s a lefordított szövegek azonnal az adatbázisba kerüljenek.

Ugyanakkor vannak olyan megközelítések (pl. a MultiTrans[26]), amelyek nem igénylik adatbázis építését, ehelyett a számítógépen vagy a hálózaton végzett teljes szövegű kereséssel és valós idejű szövegszinkronizálással találják meg a korábban már tárolt forrásszöveget és fordítását. Ilyen értelemben az asztali és internetes keresőrendszerek is fordítási erőforrásként használhatók. Sőt, vannak olyan keresőrendszerek is, amelyek különböző típusú fordítómemóriákban tesznek lehetővé együttes keresést (APSIC XBench[27]).

A fordítási egység. A fordítómemória megfelelő működéséhez a szöveget egynemű egységekre kell osztani (homogén felosztás). Olyan egységeket kell találni, amelyek határait számítógéppel könnyen meg lehet találni – ez a minimális informatikai megoldás gazdasági igényéből következik –, de amelyek még ismétlődhetnek annyira, hogy az ismétlődés kihasználása valódi hatékonyságnövekedést jelentsen.

Korábban említettük, hogy a gépi fordítástámogatás tárgya elsősorban a szakmai szövegek fordítása. A szakmai szövegek elsődlegesen és kötelezően ismétlődő része a terminológia. A terminológia azonban a szakmai szövegek tömegének kisebb részét teszi, a szövegek diskurzusában és frazeológiai bázisában a terminológián kívüli nyelvi elemek is nagy tömegben találhatók. A szöveget tehát terminusokra nem lehet felosztani, ráadásul a szöveg többi részéről való gépi leválasztásuk sem egyszerű.

Megfigyelhetjük azonban, hogy a szakmai szövegekre a terminológiai tartalmon kívüli ismétlődés is jellemző. Ez azt jelenti, hogy a Kis Ádám által terminológiai magatartásnak nevezett szövegszervezési elv a szakmai szö-

veg egészére kiterjed (Kis Á. 2003). Ennek alátámasztását szolgálja az 1. táblázat, amely két informatikai szakkönyv ismétlődésstatisztikáját hasonlítja össze egy szépirodalmi mű szövegével:

Szöveg	*Szegmentumok száma*	*Ismétlődő szegmentumok száma*	*Ismétlődés tömege a szövegszavak %-ában*
1. informatikai könyv[28]	10 228	3 883	13%
2. informatikai könyv[29]	17 315	1 581	4%
Orwell: 1984	6 553	92	0,003%

4.1. táblázat. Két informatikai szakkönyv és George Orwell 1984. c. regényének ismétlődési statisztikája, a MemoQ fordítási környezetben kiszámítva. A számítás a FNy szövegen történt

A 4.1. táblázat bevezeti a szegmentum fogalmát: ez az az absztrakt egység, amelyre a szöveget fel kell osztani ahhoz, hogy fordítómemória-adatbázisban tárolható legyen. E szegmentum általában a mondat közelítése, mivel ez az a szónál nagyobb, de még viszonylag jól ismétlődő szövegegység, amelynek határai egyszerűen felismerhetők számítógéppel.

A fordítómemóriákban tehát a fordítás egysége a mondatot közelítő szegmentum. Ez a latin, cirill, görög stb. betűkkel, általában a betűkkel írt szövegekre igaz. A szegmentálást a számítógép általában automatikusan végzi, két okból: egyfelől a felhasználó megkímélhető attól, hogy minden esetben manuálisan jeleznie kelljen azt a szegmentumot, amelynek fordítására készül, másrészt a szövegeken a kötegelt üzemmódú (emberi beavatkozás nélkül zajló) ismétlődésstatisztika-számítás és előfordítás sem végezhető el automatikus szegmentálás nélkül.

A latin betűs szövegek gépi mondatszegmentálása általában egyszerű, bár nem olyan triviális, mint a szószegmentálás, ahol a szóköz és a központozási jelek mindig szóhatárt jelentenek (más kérdés, hogy a szó mennyiben lexikai egysége a szövegnek). A mondatzáró írásjel (., !, ?) az ilyen szövegekben rendszerint valóban mondatvéget jelez. Általában a {mondatzáró írásjel, szóköz, nagybetű} sorozatot mondatvégként kezelhetjük, akárcsak legtöbbször a bekezdésvéget. Sokszor azonban a kettőspont is mondatvéget jelez, máskor a pont sem (például rövidítések vagy sorszámok végén). Ezek mindenesetre egyszerűen leírható szabályok, a megfelelő – bár nem tökéletes – mondatszegmentálás leírására forrásnyelvenként 1-3 ún. reguláris kifejezésre van szükség (Friedl 2003). Ezzel 97-98%-os szegmentálási pontosságot el lehet érni (Véronis 1998), amely jól formált szakmai szövegek esetén érvényes, és abban az esetben megfelelő, ha utólag van mód az automatikus szegmentálás javítására, legkésőbb a fordítás vagy a szövegszinkronizálás során.

A fordítómemóriák a szegmentumokat fordítási egységekbe szervezik. A fordítási egység *(translation unit,* TU) egy forrásnyelvi szegmentum (a legtöbbször egy mondat) és CNy megfelelője által alkotott pár. A fordítómemória felügyeletével végzett fordítás során ilyen fordítási egységek jönnek létre.

A fordítómemóriák hatékonysága

A hatékonyság mérése. A fordítási munka hatékonyságát általában nem mérjük, de ha mérnénk, valamilyen szövegegység lefordítására fordított időt értenénk alatta. Itt is szükségünk van egy jóhiszeműségi hipotézisre: feltételezzük, hogy a fordításra fordított idő az ekvivalens CNy szöveg előállításához szükséges, anélkül hogy az ekvivalencia fogalmát közelebbről meghatároznánk.

Saját tapasztalatom szerint a fordítók segítség nélkül legfeljebb 20 szabványoldalt (25 000 leütés, angol forrásnyelv esetén kb. 8 000 szövegszó) képesek lefordítani egy 8 órás munkanap alatt, de ez jelentősen függ a szövegtől és a fordító felkészültségétől, gyakorlatától is.

A fordítómemóriák ismerete nélkül meghökkentő lenne úgy definiálni a hatékonyságot, mint az adott dokumentumban ténylegesen lefordítandó szöveg tömegét. Ha a korpusznyelvészetben szokott módon szövegszavakban mérjük a terjedelmet, ez a következőt jelenti (jelölések tőlem):

$$\eta = 1 - \frac{w_f}{w_t}$$

ahol η a hatékonyság mértéke, w_f a lefordítandó szegmentumok együttes terjedelme szövegszóban, a w_t pedig a forrásszöveg teljes terjedelme ugyancsak szövegszóban. A hatékonyság mértéke 0, ha minden szövegszót le kell fordítani, és 0,2, ha a szövegszavak 20%-ára valamilyen automatikusan képzett javaslatot kaptunk, így csak a fennmaradó részt kell lefordítani. A fordítómemóriák gyártói ezt a számot adják meg, amikor a legfeljebb 30%-os hatékonyságnövekedést említik.

Ezt a számot mindaddig nagyon könnyű kiszámítani, amíg csak a teljes szegmentumegyezéseket vesszük alapul, illetve csak a teljes ismétlődéseket számoljuk meg. Azonban láthattuk, hogy ezen az alapon – informatikai szakszövegekben – az elérhető hatékonyság 0,05–0,09, ami nem éri el a gyártók által idézett felmérések számait.

Ha csak a teljes egyezést tekintjük (belső vagy külső) ismétlődésnek, figyelmen kívül hagyjuk például a következőt:

korábban lefordított FNy szegmentum	‘Because site mappings are independent from logical domain structures, there’s no necessary relationship between a network’s physical structure and its logical domain structure.’
aktuális FNy szegmentum	‘Sites mappings are independent from logical domain structures, and because of this there’s no necessary relationship between a network’s physical structure and its logical domain structure.’

A két mondat szókincse szinte teljesen megegyezik, csak a grammatikai kivitelezés más, vagyis a két mondat erősen hasonlít. Ha az ilyen tárolt fordításokat is visszakapjuk, a hatékonyság jobban növelhető, mint ahogy az a teljes egyezésre épülő ismétlődésekből következne.

A fenti hatékonyságértékek a szakszövegek belső ismétlődésstatisztikájára épülnek. A külső ismétlődések felhasználásával ez a hatékonyság tovább növelhető, azonban megfigyelésem alapján a szövegközi teljes egyezés rendkívül ritka (a fent idézett majdnem-egyezés inkább jellemző). Másfelől külső egyezésekre csak akkor számíthatunk, ha valóban megfelelő módon – a fordítómemória-adatbázisban – rendelkezésünkre állnak korábbi fordítások.

A hasonlósági keresés és az utószerkesztés. A fordítómemória csak úgy tudja nyújtani a gyártók által ígért maximális 0,3 értékű hatékonyságot, hogy a korábbi fordításokból nemcsak aktuális forrásszöveg szegmentumaival egyező, hanem az azokhoz hasonló szegmentumokat (és fordításaikat) is kikeresi.

A cél ekkor nem elsősorban a hasonló tartalmú szegmentumok megtalálása (az inkább asztali és internetes információkereső rendszerekben követelmény), hanem a hasonlóan fordítandó szegmentumok feltárása. Ez a követelmény ugyanakkor ebben a formában nem teljesíthető, mert az aktuális forrásszegmentum fordítása a kereséskor még nem áll rendelkezésre, ezért nem lehet a fordításokat összehasonlítani. Azonban, ha elfogadjuk az fordítói átváltási műveletek alapjául szolgáló elvet (Klaudy 1999), miszerint a fordító a fordítás során kötődik a forrásszöveg formai megvalósításához, a kereséskor akkor járunk jól, ha az aktuális forrásszegmentumhoz szókincsben és grammatikában hasonló tárolt szegmentumot keresünk. Másképp fogalmazva: olyan tárolt szegmentumot – lehetőleg egyet – keresünk, amely az aktuális forrásszegmentum lehető legkisebb mértékű variációjának tekinthető.

Hagyományosan a fordítómemória-fejlesztés egyetlen, de nem kizárólagos kutatási területe a forrásszegmentumok hasonlósági keresése, amelyre számos különböző megoldás született. Most a konkrét megvalósítástól elvonatkoztatva, a hatékonyság szemszögéből kell még tennünk néhány megállapítást.

Ha az aktuális forrásszegmentumhoz hasonlót találunk csak a fordítómemóriában, nyilván annak a fordítása is csak hasonló lesz a forrásszegmentumunk kívánt fordításához. A hasonlóság (vagy a különbség) – az átváltási műveletek elve alapján – valamilyen értelemben analóg a tárolt és az aktuális forrásszegmentum hasonlóságával (különbségével), de a felkínált és az elvárt fordítás mindenképpen különbözik. Ez azt jelenti, hogy a fordítónak a felkínált fordítást még javítania kell, ezért várhatóan nem 0 vagy minimális az a munkaráfordítás, amelyre a hasonlósági keresés mellett a forrásszegmentum fordításához szükség van.

Emiatt fontos, hogy a hasonlósági keresés találatait a fordítómemória oly módon értékelje (pontozza), hogy a pontszám (amely jellemzően 0 és 1 közé esik, és általában százalékosan adják meg) tükrözze a felkínált fordítás teljességét, vagyis közelítőleg kiszámítható legyen ennek alapján az a munkamennyiség, amennyi az elvárt fordítás kialakításához szükséges.

Példa:

korábbi FNy szegmentum	Because site mappings are independent from logical domain structures, there's no necessary relationship between a network's physical structure and its logical domain structure.
korábbi CNy megfelelő	Mivel a telephely-hozzárendelések függetlenek a logikai tartomány-struktúrától, nincs feltétlenül kapcsolat egy hálózat fizikai struktúrája és logikai tartománystruktúrája között.
aktuális FNy szegmentum	Sites mappings are independent from logical domain structures, and because of this there's no necessary relationship between a network's physical structure and its logical domain structure.
elvárt CNy megfelelő	~~Mivel~~ A telephely-hozzárendelések függetlenek a logikai tartomány-struktúrától, ezért nincs feltétlenül kapcsolat egy hálózat fizikai struktúrája és logikai tartománystruktúrája között.

A javítás két szövegszót érint, ez a tárolt fordítás 18 szövegszavának 11%-a. Ezért azt várnánk, hogy a fenti találatra a fordítómemória 89% körüli pontszámot adjon (a valós pontszám, amelyet a létező rendszer végül visszaadott, 88% volt). A pontszám kialakításának nehézsége, hogy a forrásszegmentumok, nem pedig a fordítások különbségéből kell kiszámítani.

Fontos megjegyezni, hogy a szerkesztési távolság itt használt mérőszáma jelentős egyszerűsítés eredménye, azonban belátható, hogy nem tér el jelentősen a 4.2. fejezetben leírt, módosított Levenstejn-algoritmussal számítható szerkesztési távolságtól *(edit distance).* Szerkesztési távolságnak a továbbiakban is azon szövegszavak számát tekintem, amelyeket a felajánlott CNy szegmentumban törölni vagy beszúrni kell.

A fentiek alapján módosítanunk kell a korábban felírt, egyszerűsített hatékonyságszámító képletet. Csak teljes egyezéseket figyelembe véve, de szegmentumokra bontva így írhatjuk fel:

$$\eta_e = 1 - \frac{\sum_{i=1}^{n} w_{f,i}}{\sum_{i=1}^{n} w_{t,i}}$$

ahol η_e a teljes egyezések alapján számított hatékonyság, n a forrásszöveg szegmentumainak száma, $w_{t,i}$ az i. szegmentumban lefordítandó szövegszavak száma, $w_{t,i}$ pedig az i. szegmentum szövegszavainak teljes száma. Vegyük észre, hogy ez a képlet alkalmas a hasonlósági találatok hatékonyságnövelő hatásának figyelembe vételére is. Csak teljes egyezések beszámítása esetén ugyanis csak két eset lehetséges:

$$w_{f,i} = 0 \quad \text{vagy} \quad w_{f,i} = w_{t,i}$$

Ha azonban a rendelkezésünkre áll a fordítómemória által adott pontszám, és feltételezzük, hogy az jól közelíti a tárolt és az elvárt fordítás közötti különbsé-

get, akkor a mondatonként lefordítandó szövegszavak számát a szegmentum szavainak számából és a pontszámból is kiszámíthatjuk:

$$w_{f,i} = (1 - \sigma_i) w_{t,i},$$

ahol σ_i a fordítómemória által az *i.* szegmentumra adott pontszám. Ez teljes egyezés esetén ($\sigma_i = 1$) 0, ami erős egyszerűsítés, ugyanis egyfelől nincs garancia arra, hogy a teljes egyezésből kapott tárolt fordítás teljesen megfelelő az aktuális helyen, másfelől pedig a kész fordítás beszúrása is munkával jár a fordító számára. Az előbbit a fordítómemória összetétele befolyásolja: minél heterogénebb az adatbázis (a forrásszövegek tekintetében), ez az eset annál valószínűbb. A második tényezőt az a fordítási környezet határozza meg, amely a fordítómemóriát mint erőforrást működteti.

A fentiekből következő többletmunkát a megrendelők általában elfogadják, ezért a teljes egyezésből kapott fordítások díjazása nem nulla. Ezért az előbbiekben felírt képletek nem határozzák meg pontosan a fordítás hatékonyságnövekedését (a fordítómemória-használat hatékonyságát); szükség lenne valamiféle korrekciós tényezőre. Azonban az értekezés megírásáig nem történtek nyilvánosan dokumentált mérések arra nézve, hogy a fordítás hatékonysága hogyan alakul a különböző típusú fordítómemória-találatok esetén.

A fordítóirodák és a fordítók spekulatív tényezőkkel azt számítják ki, hogy adott terjedelmű szöveg adott fordítómemóriával való lefordítása effektíve hány szövegszó lefordítását jelenti. Ehhez a fordítómemóriától kapott javaslatokat a találat minősége (a fordítómemóriától kapott pontszám) szerint kategóriákba sorolják, és minden kategóriához hozzárendelnek egy szorzót. Elvégzik a FNy szöveg analízisét a fordítómemória segítségével (erre minden fordítási környezet lehetőséget nyújt), és meghatározzák, hogy az egyes minőségi kategóriákba hány szövegszó esik. Ezeknek aztán meghatározzák a súlyozott összegét. Példa egy könyv adataira, ha csak a belső – részleges és teljes – ismétlődéseket (a szöveg homogeneitását) vesszük figyelembe:

Találati kategória	*Szorzó*	*Szövegszavak száma*	*Effektív szószám*
100%	0,3	5 105	1 532
95%-99%	0,4	1 565	626
85%-94%	0,5	2 711	1 356
75%-84%	0,7	9 460	6 622
50%-74%	1,0	48 204	48 204
Nincs találat	1,0	58 277	58 277
Összesen		125 322	116 617

Végezetül egy terminológiai megjegyzés: a kereskedelmi rendszerek a teljes egyezésen alapuló találatot pontos találatnak *(exact match)*, míg a hasonlósági keresésből származót *fuzzy* találatnak *(fuzzy match)* nevezik. Ez az elnevezés a *fuzzy* logika alapeleméből, a *fuzzy* számból következik. A *fuzzy*

szám olyan logikai változó, amelynek nemcsak két értéke (igaz vagy hamis, 0 vagy 1) lehet, hanem a kettő között tetszőlegesen sok, vagyis azt határozza meg, hogy a vele jellemzett állítás milyen mértékben igaz – a fordítómemóriákra vonatkoztatva pedig azt, hogy a tárolt szegmentum milyen mértékben egyezik az aktuális forrásszegmentummal. A hasonlósági találatok előállítására szolgáló szoftvermodulok (az ún. *fuzzy indexek*) ugyanakkor nem alkalmazásai a fuzzy logikának. (Bővebben lásd: Navarro 2001, Navarro et al. 2001., Planas 2000)

A karaktersorozat alapú fordítómemória korlátai

Ismételjük meg a hasonlóságra korábban hozott példát:

korábban lefordított FNy szegmentum	'Because site mappings are independent from logical domain structures, there's no necessary relationship between a network's physical structure and its logical domain structure.'
aktuális FNy szegmentum	'Sites mappings are independent from logical domain structures, and because of this there's no necessary relationship between a network's physical structure and its logical domain structure.'

A tárolt és az aktuális forrásszegmentum tartalmas szókincsében nem, csak grammatikai megvalósításában különbözik. Általános elvárásként fogalmaztuk meg, hogy a tárolt forrásszegmentum lehetőleg minimális variációja legyen az aktuális forrásszegmentumnak. Ez a következő hasonlósági kritériumokat jelentheti:

- a tárolt és az aktuális szegmentum szókincsének közös része haladjon meg egy küszöbértéket;
- a nyelvtani struktúrájuk legyen analóg.

A második feltétel kiértékelése lehetetlen a forrásnyelv grammatikai struktúráinak ismerete nélkül, vagyis ha azt számítógéppel akarjuk kiértékelni, szükséges mind a tárolt, mind az aktuális forrásszegmentum szintaktikai elemzése. Ez, ha nyelvfüggetlen, minimális megoldásra törekszünk, nem járható út.

A szókincs kiértékelése ugyanakkor többé-kevésbé megvalósítható nyelvi tudás nélkül, ám itt is előfordulhat, hogy ugyanaz a szó más toldalékolt formában jelenik meg a tárolt, mint az aktuális forrásszegmentumban. Ez esetben csak akkor hasonlíthatók össze, ha feltételezzük, hogy a különbözőképp toldalékolt formák, illetve a szótári alak és a toldalékolt formák karaktersorozatai kellőképpen hasonlítanak egymáshoz – ami tőváltozások, rendhagyó alakok esetén már nem igaz, úgyhogy lehetnek olyan esetek vagy forrásnyelvek, amikor a szókincs gépi összehasonlítása nem lehetséges megbízható módon, ha nem áll rendelkezésre a forrásnyelvhez szóelemző (morfológiai elemző és/vagy lemmatizáló) program.

A számítógép számára a természetes nyelvi szöveg atomi építőeleme a karakter (most figyelmen kívül hagyjuk a karakterkódolást, amely persze a karakterek további dekompozícióját is jelenti – vö. Prószéky-Kis 1999:27). Ezért ka-

rakterekben mindenképpen mérhető két karaktersorozat eltérése. Viszonylag könnyen meghatározható elem az ortográfiai szó (token) is, amelyeket a szegmentum szövegének a szóközök és a központozási jelek mentén végzett szegmentálással különíthetünk el egymástól.

Mivel a számítógép elemi művelettel csak két adatelem teljes egyezését tudja megállapítani – vagyis ha nem egyeznek teljesen, akkor a gép számára nincs semmi közös bennük –, két karaktersorozat (szegmentum) összehasonlítása nemtriviális számítógépes feladat. Ennek algoritmusait itt nem ismertetem, mivel ezek, matematikai bázisukkal együtt, évtizedek óta léteznek. A 4.2. fejezetben említettem, hogy két karaktersorozat szisztematikus összehasonlítását a leginkább felhasználható módon Levenstejn írta le (Levenstejn 1965). Az általa leírt algoritmus bármely két karaktersorozathoz hozzá tud rendelni egy számot, amely jellemzi a két karaktersorozat távolságát (különbözőségét).

Ez azonban két karaktersorozat összehasonlítására használható, a fordítómemória esetén pedig egy aktuális forrásszegmentumhoz kell kikeresni a leghasonlóbbat egy olyan adatbázisból, amely akár több százezer szegmentumot is tartalmazhat. A Levenstejn-algoritmus használata esetén az aktuális forrásszegmentumot mindig össze kell hasonlítani az összes tárolt forrásszegmentummal, amely a mai számítógépeken is rendkívül lassú, gazdaságtalan művelet, vagyis egy forrásszegmentum kikeresése tovább tarthat, mint manuális lefordítása.

Adatbázisokban hatékonyan keresni az adatbázis tartalmához rendelt keresési struktúrában, az ún. *indexben* lehet (Knuth 1994 (1973):III.442-527, Harris-Ross 2006:257-443). Az ilyen keresési struktúrák is csak azt teszik lehetővé, hogy egy vagy több kijelölt adatbázis-mező (az ún. *kulcs*) tartalmát gyorsan, szöveges mezők esetén a keresőszó vagy a szövegmező (nem pedig a teljes adatbázis) hosszával arányos idő alatt ki lehet keresni, s ennek alapján megkapjuk a teljes adatbázisrekordot. Ezek az indexek a legtöbb esetben teljes egyezésre épülnek, a hasonlósági index általában bonyolultabb struktúra.

A hasonlósági indexek jellemzően a keresendő karaktersorozatok dekompozíciójára építenek: a szegmentumokat tároláskor szókettesekre vagy szóhármasokra bontják; egy indexben ezek kereshetők. Kereséskor az új aktuális forrásszegmentumot is szókettesekre vagy szóhármasokra bontják, és ezeket egyenként keresik az indexben. A találatok közül pedig kiválasztják azt a tárolt szegmentumot, amely a legtöbb szókettesben vagy szóhármasban egyezik az aktuális forrásszegmentummal. Mindeközben általában nem az egyes szavak teljes egyezésére építenek, hanem azok hasonlóságát is figyelembe veszik. A hasonlósági indexelési eljárásokat szintén leírták a szakirodalomban (Navarro 2001, Navarro et al. 2001., Planas 2000).

A hasonlósági pontszámot az egyes szavak súlyozásával és az egyező szókettesek, szóhármasok számából számítják ki. Ezek az eljárások – megfelelő implementáció esetén – képesek arra, hogy olyan adekvát pontszámot adjanak, amely valóban korrelációban áll a tárolt fordítás utólagos javításához szükséges munkával:

If you later create a new child domain or a root domain in a new tree, the first domain controller in the new domain is assigned operations master roles automatically as well.	0,90	If you later create a new child domain or a root domain in a new tree, the first domain controller in the new domain is automatically assigned operations master roles as well.
Any domain controller hosting a global catalog should be well connected to the network and to domain controllers acting as infrastructure masters.	0,75	Domain controllers hosting the global catalog should be well connected to domain controllers acting as infrastructure masters.
For example, OUs associated with tech.la.cpandl.com contain objects for this domain only.	0,61	For example, organizational units associated with seattle.microsoft.com contain objects for this domain only.
You can set these properties for user accounts as discussed later in the chapter.	0,57	Once the account is created, you can set advanced properties for the account as discussed later in the chapter.
If this happens, you'll find that the printer freezes or doesn't send jobs to the print device.	0,39	Symptoms include a frozen printer or one that doesn't send jobs to the print device.

4.2. táblázat: Példa a karaktersorozat-alapú hasonlósági keresés által adott pontszámokra

A karaktersorozat alapú fordítómemóriák hiányossága, hogy csak teljes szegmentumok hasonlóságát vizsgálják, így a fordítás közben a tárolt szegmentumok igen nagy hányada rejtve marad – minél hosszabb az aktuális forrásszegmentum, annál kisebb a valószínűsége annak, hogy az adatbázisban lesz kellőképp hasonló tárolt szegmentum.

A fordítómemóriában így lehetnek olyan szegmentumok, amelyek egyik vagy másik része teljesen egyezik az aktuális forrásszegmentum egy részével (vagy éppen egészével), s lehetnek olyanok is, amelyek szókincsükben különböznek, de analóg szintaxist mutatnak.

Erre a problémára jelenleg két megoldás létezik, az egyik hozzáférhető kereskedelmi forgalomban, a másik egyelőre kísérleti fázisban van:

(1) Töredékkeresés karaktersorozat-alapon. A fordítómemória-rendszer megpróbálja az aktuális forrásszegmentum fordítását a fordítómemóriában levő töredékekből összeállítani. Ekkor olyan tárolt szegmentumokat keres, amelyek hasonlóak az aktuális forrásszegmentum egyes részeihez (vagy egyeznek vele). A rendszer meghatározott algoritmus szerint (például balról jobbra haladva, mindig a lehető legnagyobb részt lefedve) igyekszik a teljes forrásszegmentumot lefedni ilyen részleges találatokkal, s a fordítást a tárolt fordítások konkatenálásával kialakítani. Az ilyen találatok pontszáma nehezen számítható, és az is nehezen jósolható meg, hogy menynyi munkát igényel a felkínált „fordítás" kiigazítása.

Hátránya még ennek a módszernek, hogy csak abban az esetben működőképes, ha sok rövid forrásszegmentum van a fordítómemóriában.

(2) Nyelvi dekompozíció: a rendszer a forrásszegmentumokat nyelvi elemzéssel dekomponálja, s a dekompozíciót az új fordítások tárolásakor is végrehajtja. Ilyenformán a töredékek meghatározása grammatikailag adekvát módon történik, és ha a fordítás összeállításánál a rendszer hasonlóképpen (legalább részlegesen) figyelembe veszi a célnyelv követelményeit, akkor a fordítás kiigazításának munkája még kevesebb is lehet, mint ahogy az az aktuális forrásszegmentum és a hipotetikus tárolt szegmentum hasonlóságából következne (a pontszám kiszámítására egyelőre itt sem létezik egzakt módszer). Ezt a megközelítést ismerteti a következő rész.

A nyelvi támogatású fordítómemória

Korábban említettem, hogy minél hosszabb egy forrásszegmentum, annál kisebb a valószínűsége, hogy a fordítómemóriában találunk hozzá kellő mértékben hasonló forrásszegmentumot. Ha viszont a forrásszegmentumokat mind a fordítómemóriában való tárolásuk, mind pedig keresésük során dekomponáljuk, és a felkínált fordítást a részszegmentumokból állítjuk össze, jelentősen növeljük annak valószínűségét, hogy

a) a fordítómemóriába bevitt hosszú forrásszegmentum hasznosul (vagyis nem holt teherként tároljuk az adatbázisban), és
b) a fordítás közben előforduló hosszú szegmentumokra kapunk választ.

A kérdés csak az, mi legyen a dekomponálás vezérelve. A *fuzzy* indexelésnél és a statisztikai alapú gépi fordításnál a szöveget szókettesekre vagy szóhármasokra bontják, függetlenül attól, hogy a szókettes vagy szóhármas átnyúlik-e valamely szegmentumon belüli struktúrahatáron. A karaktersorozat-alapú fordítómemóriákban pedig csak úgy lehet megvalósítani a töredékkezelést (ott ez igazából a forrásszegmentum dekompozíciója), hogy pusztán mintaillesztéssel állítjuk össze a fordítási javaslatot. Ráadásul ott a fordítómemóriába vitt szegmentumok nincsenek dekomponálva, ezért annak valószínűsége, hogy egy hosszú tárolt szegmentum találatként újra előkerül, rendkívül kicsi.

Az alábbiakban ismertetek egy módszert, amelynek kidolgozásában magam is részt vettem (Hodász et al. 2005, Hodász G. et al. 2004). E módszer – és a rá épülő fordítómemória-modul – a forrás- és a célszegmentumok sekély szintaktikai dekomponálására épül. Alapötletét a MorphoLogic műhelyében kifejlesztett mintaalapú gépifordító-rendszer adta, de sem a végső elképzelés, sem az implementáció nem alkalmaz szorosan vett gépi fordítási műveleteket. Az alább leírt módszereket egyelőre csak az angol-magyar nyelvpár esetére implementáltuk és próbáltuk ki.

A fordítómemória alapműveletei. Egy fordítómemóriának két alapművelet kell elvégeznie:

(1) az aktuális forrásszegmentum megkeresése és fordítás felkínálása;
(2) a felhasználó által jóváhagyott forrás-cél szegmentumpár (fordítási egység) bevitele a fordítómemória-adatbázisba.

A két művelet nem végezhető el egymástól függetlenül, mert a bevitt fordítási egység forrásszegmentumának megkereshetőnek kell lennie. Ez a hasonlósági *(fuzzy)* index esetén például azt jelenti, hogy amennyiben a fordítási egységek forrásszegmentumai szókettesekre bontva kerülnek az indexbe, az aktuális forrásszegmentumot is szókettesenként kell fellapozni.

Az aktuális forrásszegmentum lefordítása. A nyelvi támogatású fordítómemória ennek során a következő műveletsort hajthatja végre (most ki-

hagyjuk azt a lépést, amelynek során teljes egyezést keres – ugyanis az implementált változatban ez megelőzi a nyelvi illesztést):

(1) A forrásszegmentum nyelvi elemzése: lemmatizálás, morfológiai elemzés és sekély szintaktikai elemzés

(2) A forrásszegmentum alapvető építőelemeinek meghatározása, a kisebb elemek fordításainak megkeresése a fordítómemóriában

(3) A fordítás összeállítása egy ún. mondatváz felhasználásával, szükség esetén egyes szavak morfoszintaktikai jegyeinek megváltoztatásával, illetve beállításával.

A mondatváz olyan minta, amelyben a (2) pontban megtalált kisebb építőelemek egyetlen absztrakt szimbólummal vannak helyettesítve. A minta ezeken kívül a mondat azon részeit tartalmazza, amelyek az elemzési algoritmus alapján nem képezhetik a kisebb építőelemek részét. Példa:

Fordítandó mondat	*'Microsoft Windows 2000 makes it possible to configure hard disk drives in a variety of ways.'*
Kisebb építőelemek	*'Microsoft Windows 2000'; 'hard disk drives'; 'a variety of ways'.*
Mondatváz (lemmatizálva)	*'[01] make it possible to configure [02] in [03].'*

Az eredeti elképzelés szerint ez rekurzív művelet lett volna, amelyben a kisebb építőelemeket ugyane folyamatnak vetettük volna alá. Ezt azonban teljesítményproblémák miatt elhagytuk, és ehelyett a kisebb építőelemek maximumát keressük a szegmentumokban.

A felajánlott fordításban maradhatnak hiányok, kihagyások, a művelet ettől még sikerrel befejezhető. Ilyen kihagyások akkor keletkeznek, ha a valamelyik kisebb építőelem vagy éppen a mondatváz fordítása nem található meg a fordítómemóriában.

Megtehetnénk, hogy az ilyen hiányokat gépifordító-rendszer segítségével pótoljuk, ehhez azonban biztosnak kell lennünk abban, hogy a gépifordító-rendszer legalább a kisebb építőelemeket megbízhatóan fordítja (erre nézve a dolgozat írásakor még nem történtek mérések).

Azonban megállapíthatjuk, hogy a kihagyásokat tartalmazó fordítási javaslat megfelelő válasz a rendszertől. Mi történik, ha ez az algoritmus nincs a rendszerben? A lehetőségek:

(1) Nincs semmilyen fordítási javaslat, mert az aktuális forrásszegmentum egészében még kellően hasonló formában sincs meg a fordítómemóriában. Ebben az esetben marad a manuális fordítás.

(2) A teljes mondatot megkíséreljük gépifordító-rendszerrel lefordítani, amelyről tudjuk, hogy az esetek többségében nem ad publikálható fordítást.

Megfigyelésünk szerint a manuális fordítás egyelőre kevesebb munkával jár, mint a hibás gépi fordítás kijavítása. A dekompozíció–kompozíció útján kelet-

kező fordítási javaslat viszont alapvetően emberi fordításokra épül, így hiányaival együtt is alkalmas a továbbjavításra.

Mivel a fentebb említett kisebb építőelemeket eddig meglehetősen absztrakt módon kezeltük, a későbbiekben ezek konkrétabb tárgyalására is sor kerül.

Új fordítási egység felvétele a fordítómemóriába. Bizonyos értelemben az új fordítási egység (forrás-cél szegmentumpár) felvétele teljesen független a forrásszöveg lefordításától. Nem tételezhetjük fel, hogy az emberi fordító által jóváhagyott fordítás bármiféle kapcsolatban van a fordítási erőforrások által esetleg korábban felajánlott komponált vagy teljesen gépi eredetű fordításokkal. Bár vannak módszerek a felhasználó tevékenységének követésére, célszerűbb feltételezni, hogy a jóváhagyott fordítási egység teljesen új.

Ez azért fontos, mert a tárolás során a forrásszegmentumot kapcsolatba kell hozni a célszegmentummal, mert a kisebb építőelemekkel saját fordításukat kell tárolni – vagyis a kisebb építőelemek fordítását ki kell emelni a célszegmentumból (amennyiben ez lehetséges). Ha a számítógép ismerné azt a folyamatot, amelynek során a fordítás keletkezett, lehetősége volna annak megismétlésével pontosan meghatározni az egyes forrásoldali kisebb építőelemek céloldali megfelelőit. Erre azonban nincs lehetőség.

Ezért az általunk kidolgozott fordítómemória-rendszer a következő eljárást hajtja végre:

(1) Elvégzi mind a forrás-, mind a célszegmentum nyelvi elemzését, mindkét oldalon meghatározva a kisebb építőelemeket és a mondatvázat.

(2) Szinkronizálja egymással a forrás- és céloldali kisebb építőelemeket. Ez sok szempontból a mondatok szinkronizálásához hasonlóan történik (Pohl 2006).

Ezen a ponton rendelkezésünkre állnak a kisebb építőelemekből és a mondatvázakból létrehozott szinkronizált mintapárok. A fordítási egységet így több, kisebb fordítási egységre bontottuk fel. Ezek már entitásként tárolhatók a fordítómemóriában.

Példa:

FNy: *'He explained his rather peculiar views on machine translation to me.'*
CNy: *'Kifejtette nekem a gépi fordításról vallott meglehetősen különös nézeteit.'*

Angol kisebb építőelem	***Magyar kisebb építőelem***
he	∅
his rather peculiar views on machine translation	a gépi fordításról vallott meglehetősen különös nézetei
me	én

Mondatvázak:

Angol	***Magyar***
$<NP_1>$ $explain_{[V]}$ $<NP_2>$ $to_{[PREP]}$ $<NP_3>$	$kifejt_{[V]}$ $<NP_3>_{[DAT]}$ $<NP_2>_{[ACC]}$

Ezek a minták példaalapú fordítórendszer feltöltésére is alkalmasak, ezért végső soron elképzelhető, hogy minden fordítási javaslatot gépifordító-rendszer állítson elő. Ebben az esetben minden mintapár valójában egy fordítási szabálynak felel meg (Carl 2001; Takeda 1996). A jelenlegi implementáció azonban még nem áll az integráció e szintjén.

Nyelvi dekompozíció és szinkronizálás. Egy mondatot sokféleképpen lehet elemezni. A nyelvi támogatású fordítómemóriának viszont olyan szintaktikai elemzést kell alkalmaznia, amely megfelel az alábbi követelményeknek:

(1) Az elemzés összehasonlítható mintákat ad vissza. Az összehasonlíthatóságnak a különböző forrásnyelvi minták között és a forrás-cél párokban is érvényesnek kell lennie.

(2) A kisebb minták viszonylag jól felcserélhetők egymással a mondaton belül. Ez azt jelenti, hogy zárt struktúrának kell lenniük, amelynek megváltoztatása nem változtatja meg a mondat nagyobb struktúráját.

Az általunk kifejlesztett implementációban háromszintű mondatstruktúrát alkalmazunk, szándékosan átugorva azokat a szinteket, amelyek különben óhatatlanul megjelennének mély mondatelemzés alkalmazása esetén. E döntés nyomán mellesleg akár sekély mondatelemzés is alkalmazható, ami jelentősen csökkenti a rendszer nyelvfüggőségét – legalábbis ami az új nyelvek bevezetésével járó költséget jelenti. A három szint:

(1) szavak

(2) főnévi csoportok *(noun phrase* – NP),

(3) mondatok (pontosabban: szegmentumok, amelyek csak közelítőleg felelnek meg mondatoknak).

1. Szószintű elemzés. Tokenizálás (szavakra, illetve szóértékű terminálisokra bontás) után a rendszer minden szót automatikusan lemmatizál, és meghatározza a főbb morfoszintaktikai jegyeiket. Ezt a forrás- és a célszegmentumon is végrehajtja. Az elemzést a MorphoLogic HUMOR morfológiai elemző rendszere végzi, de nem alkalmazunk egyértelműsítő modult.

Ez a lépés egy vagy több olyan „nyelvtani" mintát ad vissza, amelyek az egyes szavak morfoszintaktikai címkéjét és lemmáját tartalmazzák. Ez a legalapvetőbb fordítási minta. Példa:

'The big dog saw two cats.'

Az elemzés eredménye:

```
the[DET]
big[ADV],big[ADJ]
dog[ADV],dog[N],dog[V]
saw[N],saw[V],see[V][PAST]
two[NUM]
cat[N][PL]+period[PUNCT]
```

2. Főnévicsoport-elemzés (NP-kivonatolás, *NP chunking*). A legtöbb mondatszerkezetben az igei szerkezetben levő argumentumok általában helyettesíthetők más, azonos grammatikai szerepű kifejezéssel. Az egyes argumentumok szókincse és belső struktúrája a legkülönbözőbb lehet, azonban a külső minta szempontjából ez érdektelen.

Az angol igei szerkezetek argumentumai általában elöljárószói csoportok (*prepositional phrase* – PP), amelyek egy elöljárószóból és egy főnévi csoportból (NP) állnak. Ezek közül az utóbbi kicserélhető, az előbbi viszont (részben) épp az NP mondatbeli szerepét határozza meg.

Megjegyzés. A főnévi csoport azért is jó választásnak bizonyult, mert a rendelkezésre álló angol-magyar gépifordító-rendszer a dolgozat írása idején a sok főnévi csoportra jó minőségű fordítást ad, s elképzelhető az, hogy a fordítómemória által felkínált fordításokban levő hiányok nagy része előbb-utóbb automatikus eljárással kipótolható lesz.

A magyarban az „elöljárószói csoport" az esetragos vagy névutós főnévi csoport alakját ölti. Az előbbi esetben az esetrag a főnévi csoport fejének toldaléka (amennyiben fejközpontú megközelítést alkalmazunk, bővebben lásd: Pollard-Sag 1995, Trón 2000).

Itt az a nehézség, hogy a „tiszta" NP-t el kell különíteni azoktól az elemektől, amelyek az igei szerkezeten belüli szerepét határozzák meg. Ehhez a feladathoz mind a forrás-, mind a célnyelvre nagy pontosságú NP-elemző szükséges. Ez lehet mély elemző NP-elemző részhalmaza, de önálló sekély elemző is.

Az NP-elemzés során nem őrizzük meg a főnévi csoportok belső szerkezetét, csak sekély struktúrát: a rendszerünk által reprezentált NP-minta morfoszintaktikai címkék és lemmák sorozata. A köztes szinteket azért hagyjuk ki, mert ezek minden NP esetén mások lehetnek, s ezért nem alkalmasak a különböző szegmentumok közötti összehasonlításra.

Példa NP-elemzésre és fordítómemória-specifikus NP-struktúrákra (a Meta-Morpho-formalizmusban):

Forrásszegmentum: *'The big dog saw two cats.'*

A rendszer által megtalált leghosszabb főnévi csoportok:

EN.NP-FULL 50 (NP 47)
DET lex=„the"
ADJ lex=„big"
N lex=„dog" num=SG
EN.NP-FULL 282 (NP 280)
NUM lex=„two"
N lex=„cat" num=PL

3. Mondatvázak. A teljes forrás- és célszegmentum morfológiai elemzése alacsony szintű mondatmintát eredményez, amelynek elemei lemmatizált szóalakok. Amikor a rendszer megállapítja a főnévi csoportok határait, az NP-knek megfelelő részmintákat egyetlen, absztrakt NP szimbólummal helyettesíti. Az előbbi példában ez a következőt jelenti:

EN.S-FULL 363
NP 47
V lex=„see” form=F2
NP 280
PUNCT lex=„period”

A mintában így NP-helyek jelennek meg, ahová gyakorlatilag tetszőleges más NP behelyettesíthető. A mondatváz így olyan minta, amelyben a funkcionális összetevők (igék, elöljárószavak és más NP-be nem tartozó szavak) továbbra is jelen vannak (lemmával és morfoszintaktikai címkével), míg a főnévi csoportok helyén NP-lyukak jelennek meg.

E lépés során a rendszer szétválasztja a mondatvázat és a felszíni főnévi csoportokat. Ha több mondatváz és NP van, ezek tetszőlegesen kombinálhatók. A rendszer így olyan forrásszegmentumokhoz is tud komponált fordításokat felajánlani, amelyek egészükben sehol sem fordultak elő a fordítómemóriába bevitt szövegekben.

Az elöljárószókat, az esetragokat és a többi, az NP-k mondatbeli szerepét meghatározó elemet megszorításként (követelményként) meg kell tartani a mondatvázban. Az NP-mintákban pedig meg kell jelölni egyes szimbólumokat és jegyeket: ezáltal NP-k kombinatív beszúrásakor azok morfoszintaktikai jegyeit a kiválasztott mondatvázhoz lehet igazítani (ennek megvalósítása a dolgozat írása idején erősen kísérleti fázisban volt).

4. Főnévi csoportok szinkronizálása. Amikor új fordítási egységet viszünk be a fordítómemóriába, a forrás- és a célszegmentumot egyaránt elemezni kell. Feltételezhetjük, hogy a forrásszegmentum minden NP-jének lesz fordítása a célszegmentumban – ez a feltételezés működik, amikor a fordítómemória-modul új fordítási javaslatot. állít össze. Azonban egyes átváltási műveletek természete miatt ezt nem feltételezhetjük, amikor az emberi fordító által jóváhagyott fordítási egységeket dolgozzuk fel. Az NP-k megfelelését ezért külön eljárással kell megállapítani.

Ehhez néhány heurisztikus eljárást lehet használni: megfeleltethetjük egymásnak az egyes NP-k mondatbeli szerepét leíró jegyeket, és alkalmazhatunk szótáras módszereket, amelyekkel a forrás- és céloldali NP-kben levő tartalmas szavak megfelelését vizsgáljuk (ez hatékonyabb, ha terminológia is rendelkezésre áll, mert annak fordítása egyértelműbb).

Nem alapkövetelmény, hogy a rendszer egy fordítási egységben minden NP-t tökéletesen szinkronizáljon. Ehelyett úgy döntöttünk, hogy a fordítómemóriába csak a sikeresen szinkronizált forrás-cél párokat vesszük fel, és kihagyjuk azokat az NP-ket, amelyekhez az algoritmus nem talált párt. Ennek az lehet az oka, hogy a fordítása nem jelenik meg a célszegmentumban (pl. a fordító nem fordítja le az angol személyes névmást), de az is, hogy az eredeti hiányos mondat fordításában megjelenik egy főnévi csoport (explicitáció).

A rendszer értékelése. Az alábbiakban azokat az érveket sorakoztatom fel, amelyek alapján nyilvánvaló, hogy a nyelvi támogatású fordítómemória hosszú szegmentumok esetén lényeges hatékonyságjavulást jelent a kizárólag karaktersorozat-alapú fordítómemóriákhoz képest.

A nyelvtechnológiai eszközök legfontosabb számszerűsíthető minőségi mutatói a fedés *(recall)* és a pontosság *(precision)*.

Ha ekkor a fedést úgy értelmezzük, mint azoknak a fordítómemóriabeli szegmentumoknak az arányát, amelyeknek 50%-nál magasabb esélyük van arra, hogy találatban megjelenjenek, akkor már világossá válik a nyelvi alapú fordítómemória előnye: a rövidebb forrásszegmentumok találatban való megjelenésének valószínűsége nagyobb. A nyelvi alapú fordítómemória pedig rövidebb szegmentumokat tárol, mint a matematikai alapú, mivel minden esetben felbontja a mondatokat főnévi csoportokra és a főnévi csoportok helyén egy absztrakt szimbólumot tartalmazó mondatvázakra.

Ami a pontosságot illeti: egyedül a nyelvi alapú fordítómemóriának van esélye arra, hogy a fordítás utószerkesztésében segítsen. A matematikai alapú rendszer csak az egész mondat „konzervként" tárolt fordítását tudja felajánlani, míg a nyelvi alapú minden esetben egymástól független mondatvázak és főnévi csoportok kombinálásával állítja elő a fordítást. Ez már önmagában „utószerkesztés", bár nem utólag történik. A főnévi csoportokat ezután be kell illeszteni a mondatvázba, ami azt jelenti, hogy egyes szavaikat a mondatvázban szereplő absztrakt szimbólum jegyeinek megfelelően el kell ragozni (ha a célnyelv – mint jelen esetben a magyar – ragozó nyelv). Erre a célra a rendszer morfológiai generátor modult tartalmaz.

Az eddigekben a rendszer teszteléséhez két komponensből álló tesztadatokat használtunk: az egyik komponens a gyakorlókorpusz *(training corpus;* ez a nagyobb), a másik a tesztkorpusz *(test corpus;* ez a kisebb). A gyakorlókorpuszból felépítettük a nyelvtechnológiai eszköz lexikonját és/vagy szabálybázisát, majd a tesztkorpuszon – amelynek tartalmát tekintve függetlennek kell lennie a gyakorlókorpusztól – kipróbáltuk az így betanított fordítómemóriát.

Tesztünkben a gyakorlókorpusz az informatikai szövegeket tartalmazó párhuzamos SZAK korpusz volt (Kis Á.-Kis B. 2003 – a 4.2. részben leírt SZAK javítókorpusz elődje), a tesztkorpusz pedig egy informatikai tárgyú könyv szövege, amely forrását tekintve független volt a SZAK korpusztól, témáját és regiszterét tekintve azonban nem.

A konkrét tesztelés során a tesztkorpuszt alkotó szöveget szúrópróbaszerűen kipróbáltuk a nyelvi alapú fordítómemóriában, oly módon, hogy szimultán módon matematikai fordítómemóriát is használtunk, és vizsgáltuk a két rendszer találatai közötti különbséget.

fordítandó mondat	*matematikai hasonlóság*		*nyelvi hasonlóság*
	mért.	*felajánlott fordítás*	*felajánlott fordítás*
where FileName is the name of the file to which you want to write the IP configuration information.	0,41	Márta a barátjának a neve, akinek a levelét keresi.	[ahol] [FileName] [a fájlnak a neve] [amelybe ki kell írni] [az IP-konfiguráció adatát].
Getting Started with the Active Directory Command-Line Tools	0,55	Bevezetés az Active Directory-ba	[Bevezetés] [az Active Directory] [parancssori eszközbe].
You must use the ASSIGN command to do this.	0,50	Ehhez a helyreállítási konzolt (Recovery Console) kell használnunk.	[Ehhez] [az ASSIGN parancsot] [kell használnunk].
Another great resource for Windows utilities is the Microsoft Windows Server 2003 Resource Kit.	0,42	Telepíti a Microsoft Windows 2000 Resource Kit-et	[Egy másik kiváló forrás] [Windows-segédprogramok] [a Microsoft Windows Server 2003 Resource Kit].
In a domain, this means you must be a member of the Administrators, Print Operators, or Server Operators group.	0,47	A Windows 2000 kiszolgáló mappáinak megosztásához az Administrators (Rendszergazdák), vagy a Server Operators (Kiszolgálófelelősök) csoportjához kell tartoznunk.	[Tartományban][,] [ez azt jelenti, hogy] [az Administrators][,] [a Print Operators][,] [vagy] [a Server Operators csoport] [tagjának] [kell lenni].

4.3. táblázat. A karaktersorozat-alapú és a nyelvi támogatású fordítómemória találatainak összehasonlítása

A 4.3. táblázatból jól látszik, hogy azon mondatok esetén, amelyekhez a karaktersorozat-alapú fordítómemória gyengébb minőségű találatokat javasol (és amelyeket a tényleges alkalmazások fel sem ajánlanak), a nyelvi támogatású fordítómemória olyan fordítási javaslatokat állított össze, amelyek tartalmilag sokkal közelebb vannak az aktuális forrásszegmentumokhoz, mint a karaktersorozat-alapú fordítómemória javaslatai. Bár a nyelvi támogatással javasolt fordításokat is kell időnként javítani utólag, a javítás más jellegű (kisebb, kihagyott blokkok fordítása és a nyelvtani szerkezet helyreállítása). Emellett pedig a nyelvi támogatású fordítómemória jelezni tudja, hogy a javaslatot milyen blokkokból állította össze, ezért a potenciális javítási helyek azonnal látszanak – míg a karaktersorozat-alapú fordítómemória esetében szinte mindig végig kell olvasni a javaslatot. Azonban hangsúlyozzuk, hogy a matematikai fordítómemóriák – éppen ezért – a 40-50%-ra pontozott találatokat már nem adják vissza.

Mindent egybevetve kijelenthetjük, hogy a nyelvi támogatású fordítómemória a karaktersorozat-alapú fordítómemóriához képest jelentős hatékonyságnövekedést jelent, mert

(1) olyan esetekben is javasol tartalmilag adekvát fordítást, amikor a karaktersorozat-alapú fordítómemória javaslatának tartalma már nagyon távol esik az aktuális forrásszegmentumtól – vagy éppen nem is jelenik meg javaslat;

(2) az utólagos javítás igényét azáltal is csökkenti, hogy egyrészt a főnévi csoportok ragozását a mondatsablonhoz igazítja, másrészt pedig megjelöli a potenciális javítási helyeket, így a javaslat annak előzetes végigolvasása és teljes értelmezése nélkül is kijavítható. Pszicholingvisztikai szempontból ez azt jelenti, hogy a javítást a fordító vagy a lektor „sekélyebb" nyelvi műveletekkel – kis részek fordításával és kiemelt nyelvtani hibák javításával – el tudja végezni.

A fordítómemóriák értékelési szempontjai és módszerei

Értékelési szempontok. Az itt következő szempontokat magam dolgoztam ki, igazolásuk további kutatást igényel. Mindazonáltal úgy vélem, hogy a leírt mértékek és módszerek logikája magáért beszél, a leírás pedig kimutatja helyességüket.

A fordítómemóriák elsődleges rendeltetése az, hogy a fordítási munkát hatékonyabbá (külső megfigyelő számára gyorsabbá) tegyék. A fordítómemória akkor hasznos, ha minél több alkalommal talál meg forrásnyelvi szegmentumokat az adatbázisában, és a találatok is jó minőségűek. A jó minőségű találat azt jelenti, hogy a fordítómemória által ajánlott fordítás közvetlenül, illetve minimális javítással felhasználható. Önmagában az, hogy egy forrásszegmentum azonos formában megtalálható az adatbázisban, nem jelenti azt, hogy változtatás nélkül felhasználható, hiszen a szövegkörnyezettől, a kommunikációs helyzettől függően esetleg még azt is meg kell változtatni.

Megjegyezzük, hogy a szövegbeli szegmentumok – mondatok – környezetével, a fordítás környezetfüggő kiigazításával a kereskedelemben kapható fordítástámogató eszközök nem foglalkoznak, sőt az ez irányú kutatások is meglehetősen ritkák. A fordítók, fordításszervezők általában úgy segítenek ezen a problémán, hogy a szövegek témája, tárgyköre, műhelynyelve szerint külön fordítómemória-adatbázisokat készítenek és használnak. A megfelelő fordítómemória kiválasztása ezután a felhasználó dolga. Ebben kutatási feladat, hogy a számítógép – a szöveg felületes elemzése alapján – valamiféle automatizmussal segítse a megfelelő témájú, regiszterű fordítómemória-adatbázis kiválasztását.

A fentiek globális elvárások a fordítómemóriákkal szemben. Mivel azonban a fordítómemóriák jellemzően üresen kerülnek a felhasználókhoz, és mindig a felhasználás helyén töltik fel őket, a rendszer minősége – a technológiától függetlenül – nagyban függ attól, hogy milyen mennyiségű és minőségű forrásszöveg kerül bele. Ez kívülről nem szabályozható, ezért a fordítómemória-technológiák minősége globálisan nem mérhető, és egyszerű mérésekkel nem is hasonlítható össze. Megbízható összehasonlító adatokat csak majd több éves tesztelésből és fordítói statisztikából nyerhetünk.

Könnyen végrehajtható és megismételhető mérést a nyelvtechnológia statisztikai módszerei tesznek lehetővé. Ha peremfeltételként rögzítjük a fordítómemória tartalmát (a fordítómemóriába felvett korpuszt) és a lefordítandó – a fordítómemóriában nem szereplő – szöveget, akkor mérhető a nyelvtechnológiai

eszközök értékeléséhez használt két legfontosabb paraméter: a fedés *(recall)* és a pontosság *(precision)*.

A fedés egyfelől jelentheti a lefordítandó szövegből a fordítómemóriában – részben vagy egészben – megtalált szegmentumok számát, de azt is, hogy a szöveg – vagy több szöveg – lefordításakor a fordítómemóriában tárolt szegmentumoknak mekkora hányadát találtuk meg. A fordítómemóriában ugyanis elveszhetnek szegmentumok: lehetnek olyanok, amelyek a bevitelük után évekig – vagy éppen soha – nem kerülnek már találatba.

Az első definíció alapján a következő képleteket írhatjuk fel, az elsőt a szegmentumok, a másodikat a szövegszavak száma alapján:

$$r = \frac{n_h}{n},$$

ahol r a fedés, n_h azon szegmentumok száma, amelyre a fordítómemória adott találatot, n pedig a szöveg szegmentumainak teljes száma. Ez – ha a fordítómemória által adott pontszámot mérvadónak fogadjuk el – a pontszámokkal súlyozható. Ez azt jelenti, hogy a 80%-os (0,8 pontszámú) találat az adott szegmentumot 80%-ban fedi le:

$$r = \frac{\sum_{i=1}^{n} \sigma_i}{n},$$

ahol σ_i az i. szegmentumra adott legjobb találat pontszáma. $\sigma_i = 0$, ha nincs találat.

Ha a fentieket a szövegszavak számával írjuk át, talán pontosabb közelítést kapunk:

$$r = \frac{\sum_{s \in H} w_s}{\sum_{i=1}^{n} w_j},$$

ahol H azon szegmentumok halmaza, amelyekre a fordítómemória találatot adott, w_s az aktuális szegmentum, w_j pedig a j. szegmentum szövegszavainak száma.

A fordítómemória által adott találatokkal súlyozva:

$$r = \frac{\sum_{i=1}^{n} \sigma_i w_i}{\sum_{i=1}^{n} w_i},$$

ahol w_i az i. szegmentum szövegszavainak száma.

Ezt az értéket befolyásolja a szöveg belső ismétlődése, hiszen az ismétlődő szegmentumok első előfordulásuk után már találatként jelentkeznek. Így a fordítómemória és a forrásszöveg nem függetlenek egymástól, vagyis a fedést csak adott fordítómemória és forrásszöveg együttesére érdemes kiszámítani.

A másik megközelítés – amikor a fordítómemória kihasználtságát mérjük – a következőképpen írható fel képlettel:

$$r_m = \frac{n_{h,m}}{n_m} ,$$

ahol $n_{\mathrm{h,m}}$ a fordítómemória azon fordítási egységeinek száma, amelyek már részt vettek találatban, n_{m} pedig a fordítómemóriában tárolt teljes szegmentumok száma. Nagyon fontos, hogy ez nem statikus érték, mert a fordítómemória tartalma minden egyes forrásdokumentum feldolgozásakor változik. Az ilyen fedés tehát időben változó, az idő függvényében felépülő érték, amelynek időbeli változását érdemes is figyelni:

$$r_m(t) = \frac{n_{h,m}(t)}{n_m(t)}$$

A pillanatnyi érték a szegmentumok száma helyett a szövegszavak számával felírva:

$$r_m = \frac{\sum_{s \in H_m} w_{m,s}}{\sum_{i=1}^{n_m} w_{m,i}},$$

ahol $w_{\mathrm{m},s}$ a találatként visszaadott s szegmentum szövegszavainak száma, H a találatként visszaadott szegmentumok halmaza, $w_{\mathrm{m},i}$ pedig az i. szegmentum szövegszavainak száma.

A p o n t o s s á g a találatok nyomán felajánlott fordítások minősége: ennek igazi mérése azt jelentené, hogy a találatok között megszámoljuk azokat, amelyeknél nem volt szükség utójavításra, illetve minden felajánlott fordítás esetén figyelnénk, hogy annak utólagos átigazítása a fordítás szövegének hány százalékát érinti. Az utóbbiak összegének komplementere lenne a pontosság. Ennek mérését külön nem javaslom, a hatékonyság mérőszáma (pontosabban képlete) jól közelíti.

A fentiek alapján a fordítómemóriák minőségét három szempont szerint mérhetjük, illetve értékelhetjük:

(1) kihasználtság (r_{m}): a fordítómemória-adatbázis időben változó mennyiségi mutatója, amely egyfelől utal a fordítómemóriát lekérdező algoritmus hatékonyságára, másfelől pedig közvetlenül megmutatja, hogy a fordítómemóriában mekkora a „holt teher" – a tárolt, de fel nem használt szegmentumok tömege. Ennek hosszú távú, adott algoritmus mellett több fordítómemóriára kiterjedő mérése adja meg az algoritmus egyfajta jellemzését.

(2) hatékonyság (η): a fordítás hatékonysága (hatékonyságnövekedése) a korábban leírt értelemben, adott fordítómemória, fordítómemória-algoritmus és forrásszöveg mellett, a fordítás elvégzéséhez szükséges tényleges munka alapján. Közvetlenül a gépi fordítástámogatás hasznosságát méri.

(3) informativitás (ι): a fordítómemória által a találatokhoz adott pontszámok *(scoring)* értékelése, annak meghatározása, hogy az adott találathoz adott pontszám milyen korrelációban állnak az adott szegmentum kijavításához szükséges munkával. Ez közvetlenül a fordítómemóriát lekérdező algoritmus minőségét méri.

A (3) értékről az eddigiekben feltételeztük a megfelelő értéket, azonban szükséges az értékelése és a formalizálása is:

$$\iota = \mathrm{P}\left(\sigma_s \cong 1 - \frac{w_{t,f}}{w_t} - \delta \right),$$

ahol σ_s egy meghatározott forrásszegmentumra adott találat pontszáma, $w_{t,f}$ a forrásszegmentum mellett tárolt célszegmentum átigazításakor megváltoztatott (törölt, beszúrt, javított) szövegszavak száma, w_t a tárolt célszegmentum szövegszavainak száma, δ pedig a fordítástámogató eszköz által a folyamatba bevitt többletmunka korrekciós tényezője.

A következőkben mindhárom jellemző mérésére, értékelésére felvázolok egy módszert, azonban hangsúlyozom, hogy ilyen méréseket eddig nem végeztem, ezért az itt leírtak kutatási javaslatnak tekinthetők a fordítástudomány vagy a nyelvtechnológia kutatói számára. Ugyancsak rájuk marad az itt felvázolt képletek igazolása és pontosítása is.

A fordítómemória kihasználtságának mérése. Ezt akkor könnyű mérni, ha a fordítómemóriát kezelő program jegyzi, hogy mely fordítási egységek vettek addig részt találatban, s melyek nem. Amennyiben erre nincs lehetőség, egy kísérleti perióduson keresztül össze kell gyűjteni az adott fordítómemória felhasználásával fordított szövegeket, és meg kell számolni bennük azokat a szegmentumokat, amelyek a fordítómemória által adott találatból származnak.

Itt fontos, hogy csak a különböző szegmentumokat szabad figyelembe venni, ráadásul nem elég a talált szegmentumok egyezését figyelni, hiszen hasonlósági találatok több különböző szegmentum esetén is előhívhatják ugyanazt a szegmentumot a fordítómemóriából.

Programozói segítség nélkül ez a következőképpen oldható meg: feltételezzük, hogy az alkalmazott fordítómemória-programnak van előfordítás funkciója. Ha megőriztük az eredeti forrásszövegeket, a fordítástámogató rendszer segítségével előfordíthatjuk őket. Az előfordítás eredményéül kapott találatokat nem igazítjuk ki, ehelyett megszámoljuk, hogy a folyamat során a fordítómemória hány különböző szegmentumot adott vissza. Ebben az esetben ez azért releváns, mert a fordítómemória nagyon hasonló forrásszegmentumokra várhatóan ugyanazt a fordítási egységet adja vissza, annak pedig nagyon kicsi a valószínűsége, hogy két eltérő forrásszegmentumhoz betűre ugyanaz a fordítás tartozik.

Ha a fentiekből megkaptuk, hány különböző célszegmentum jelent meg az előfordítás kimenetén, közelítőleg megkaptuk tehát az előző rész képletében szereplő $n_{h,m}$ számot. (Az n_m szám – a fordítómemóriában levő fordítási egységek teljes száma – a fordítástámogató programoktól rendszerint egy lépésben megkérdezhető.) Azért közelítőleg, mert ha a fordító interaktívan dolgozik, egyes rendszerekben több találatot is kap egy forrásszegmentumra, amelyekből szabadon választhat – így az előfordítás során nem feltétlenül ugyanaz a célszegmentum kerül a szövegbe, mint a fordító interaktív munkája közben.

A fordítási hatékonyság (pontosság) mérése: Szoftverergonómiai módszerek. Ennek során két dolgot kell megmérnünk:

(1) a fordítómemória által felajánlott találatok kiigazításához szükséges munka: ezt mérhetjük időben vagy számolhatjuk, hány szót kell megváltoztatni a felajánlott célszegmentumban; ez az egyes szegmentumok esetén a w_f szám.

(2) a találatok lekérdezéséhez, a fordítás megerősítéséhez és a fordítás (kiigazítás utáni) véglegesítéséhez szükséges munka mennyisége, amelyet először időben mérhetünk, majd meghatározhatjuk, hogy a kiigazítási műveletekhez hogy viszonyul, ekkor kapjuk meg a – fordítástámogató rendszerre jellemző – δ számot.

Felvetődik, hogy miért nem lehet a hatékonyságot egyszerű kísérlettel mérni: kialakíthatnánk két fordítócsoportot, hogy az egyik fordítómemóriával, a másik pedig a nélkül fordítsa le ugyanazt a szöveget – s ezután összehasonlíthatnánk a munkára fordított időt. Ez több szempontból is nehézséget jelent:

- a fordítók képességei eltérőek, ezért ahhoz, hogy mindkét csoport átlagos teljesítményt mutasson, legalább három-öt fordítóra szükség van;
- rövid forrásszövegek esetén az eredmény nem lesz szignifikáns: olyan forrásszövegeket kell találni, amelyek – fordítómemória nélküli – lefordításához több nap szükséges.
- az eredmény csak akkor releváns, ha a fordítómemóriát alkalmazó fordítók gyakorlottan tudják kezelni a kísérlethez használt fordítástámogató programot.

Ez tehát azt jelenti, hogy a kontrollcsoportos kísérlet erőforrásigénye (költsége) adott esetben túl nagy ahhoz, hogy hatékonyan elvégezhető legyen. Emiatt lehet szükség szoftverergonómiai mérésekre, ami alapvetően a felhasználó tevékenységének valós idejű vagy utólagos követését jelenti. (Tehát nem egyszerű időmérésről van szó.) Alább leírok egy lehetséges, egyszerűen elvégezhető mérést.

A mérés peremfeltételeként rögzíteni kell a kiinduló fordítómemória-adatbázist és a lefordítandó forrásszöveget. Ezen két műveletet kell párhuzamosan elvégezni:

a) Előfordítást végezni a fordítómemóriával.
b) Interaktív módon le kell fordíttatni a szöveget egy gyakorlott fordítóval, akinek ugyanazt a fordítómemóriát kell használnia.

A két folyamat kimenetét – a célszegmentumokat – utólag össze kell vetni. A különbségképzés során az egyes szövegszavak törlését, beszúrását és megváltoztatását kell észrevenni: ehhez a 4.2. részben leírt módosított Levenstejn-algoritmus használható.

A fordítómemória találatainak értékelése. A találatok értékelése inkább a fordítómemória által visszaadott pontszámok *(scoring)* értékelését jelenti. A cél – amennyiben az eredményeket fordítómemória-algoritmus fejlesztéséhez használjuk fel –, hogy a rendszer által visszaadott pontszám erős korrelációt mutasson a felajánlott célszegmentum javításához szükséges munkamennyiséggel. Ezt manuálisan és automatikusan is el lehet végezni.

A manuális eljárás azt jelenti, hogy valamilyen kötegelt eljárással kinyerjük a fordítómemóriából az adott forrásszöveg szegmentumaira vonatkozó találatokat és pontszámokat, például ebben a formában:

Mondat a forrásszövegből	*Találat pontszáma*	*Találat a fordítómemóriában*
Before you try to restore the striped set, you should repair or replace the failed drive.	0,99	Before you try to restore the stripe set, you should repair or replace the failed drive.
These stripes are written sequentially to all drives in the striped set.	0,98	These stripes are written sequentially to all drives in the stripe set.
If your network consists of multiple physical networks, you must use multiple network adapters, with each network adapter being assigned an IP address in a different physical network segment.	0,87	If your network is divided into multiple physical networks, you must use multiple network adapters, with each network adapter being assigned an IP address in a different physical network segment.
You can assign a dynamic IP address to any of the network adapters on a computer, provided there is a DHCP server available on the network.	0,70	If the network has a DHCP server, you can assign a dynamic IP address to any of the network adapter cards on a computer.

Ezeket a pontszámokat a következő módszerek egyikével kell kiértékelni:

- manuálisan saját pontszámot rendelünk a találathoz;
- manuálisan (pl. ötös skálán) pontozzuk a pontszámot;
- valamilyen független, a fordítómemóriában feltehetőleg nem alkalmazott távolságszámítási eljárással (pl. Levenstejn 1965) új pontszámokat rendelünk a találatokhoz.

A fenti módszereket együtt is lehet alkalmazni; kiértékelésük további kutatás tárgya.

Az automatikus eljárás az előző részben leírt protokoll követése. Ugyanazt a forrásszöveget dolgozzuk fel kétszer, párhuzamosan (egyszer előfordíttatjuk, egyszer pedig interaktívan lefordíttatjuk). Meghatározzuk az előfordítás, illetve az interaktív fordítás által adott célszegmentumok különbségét, és ezt vetjük össze az egyes forrásszegmentumok találati pontszámaival. Az előző részben leírtakon túl ehhez az előfordítás találati pontszámait is ki kell nyernünk a rendszerből.

Nyelvfüggetlen módszerek a fordítómemóriák kihasználtságának javítására

A kezelőfelületen minden fordítási környezet megmutatja, hogy az aktuális FNy szegmentum milyen módon, mely szavakban tér el a fordítómemóriában talált FNy szegmentumtól, pontosabban azt, hogy az adatbázisban talált FNy szegmentumból milyen szerkesztési műveletekkel lehet eljutni az aktuális FNy szegmentumhoz (vö. Levenstejn 1964). Látjuk a törlést, a beszúrást és az átírást. A fordítómemória segítségével azt is fel lehet ismerni, hogy az adatbázisban tárolt FNy szegmentum töredéke az aktuális FNy szegmentumnak, azaz pontosan megegyezik annak valamely részével.

Ezért már a nyelvfüggetlen eljárásokkal is lehet javítani a fordítómemória kihasználtságát. Az általunk kifejlesztett MemoQ fordítási környezet az értekezés írása idején tartalmaz töredékkeresést, ami azt jelenti, hogy az aktuális FNy szegmentum kijelölt részéhez az adatbázisban meg lehet keresni azokat az FNy szegmentumokat, amelyek teljes egészükben megtalálhatók az aktuális FNy szegmentumban.

Ennek továbbfejlesztése lenne az az eljárás, amely megkísérli összeállítani a CNy szegmentum tartalmát az FNy szegmentum töredékeiből, ahol a töredékeket nemcsak az aktív fordítómemóriákban, hanem az aktív terminológiai adatbázisokban keresi. Ennek az eljárásnak a szempontjából a fordítómemória és a terminológiai adatbázis egyneműnek tekinthető; mindkettőben az aktuális FNy szegmentum rész-karaktersorozatait keressük, és mindkettőből felhasználjuk a megtalált rész-karaktersorozataihoz tartozó CNy megfelelőt. Azonban, kipróbálva a fenti, töredékkeresésre épülő eljárásokat, a tapasztalat azt mutatta, hogy a fordítási – mikrostratégiai – szituációk elenyésző hányada esetén alkalmazhatók.

Gyakrabban fordul elő, hogy az adatbázisban talált FNy szegmentum egy főnévi csoportban – terminusban – tér el az aktuális FNy szegmentumtól. Példa:

Tegyük fel, hogy az aktuális FNy szegmentum az alábbi mondatot tartalmazza:

> The most serious feature-set control problem is the problem of creeping requirements, requirements that are added late in a product's development. *(McConnell, 1996)*

Az adatbázisban pedig az alábbit találhattuk:

FNy: The most serious feature-set control problem is the problem of creeping featurism, requirements that are added late in a product's development.

CNy: A szolgáltatáskészlet szabályozásának legnagyobb problémája a lappangó szolgáltatásburjánzás; ez olyan követelményeket jelent, amelyek a fejlesztés késői szakaszában kerülnek a termékbe.

Most tegyük fel, hogy az aktív terminológiai adatbázis tartalmazza a következő két szócikket:

FNy: 'requirement', CNy: 'követelmény'
FNy: 'featurism', CNy: 'szolgáltatásburjánzás'

A fordítási környezet meg tudja állapítani, hogy a két FNy szegmentum hol tér el, és a CNy szegmentumban a módosított szó fordítása helyére be tudja írni az új szó fordítását. A toldalékolt szóalakok megtalálásához szükség lehet nyelvfüggő technológiára – lemmatizáló programra, azonban az sok nyelvhez viszonylag könnyen elérhető.

Azonban még annak is viszonylag kicsi a valószínűsége, hogy az eltérést jelentő szó vagy szócsoport adatbázisbeli változata megtalálható lesz a terminológiai adatbázisban. Ha a fordítási projekt az 5. fejezetben leírt terminológiai munkafolyamatot követi, akkor a terminológiai adatbázis az aktuális FNy szöveg terminológiai vázát képezi le. Ezért valószínűbb, hogy inkább aktuális FNy szegmentum szignifikáns főnévi csoportjait lehet megtalálni benne.

Az utóbbi esetben ismeretlen marad az adatbázisbeli FNy szövegben módosított szó fordítása, vagyis külön probléma az adatbázisban tárolt CNy szövegben megtalálni a módosuló szövegrész fordítását. Ez a kétnyelvű terminuskivonatolásban is probléma. Folytak kísérletek párhuzamos korpuszok szó-, illetve főnévicsoport-szinkronizálására (Callison-Burch et al. 2005, Choueka et al. 1994, Pohl 2006), azonban az előbbiek rendkívül nagy korpuszt, az utóbbiak pedig költséges nyelvspecifikus adatokat igényelnek, ezért a gyakorlati fordítástechnológia számára tulajdonképpen egyik sem érhető el – erről bővebben is szó lesz az 5. fejezetben, a kétnyelvű terminuskivonatolás tárgyalásánál.

Az értekezés írása idején arra folytatunk kísérleteket, hogy a fordítómemóriák konkordanciafunkciója hogyan bővíthető ki úgy, hogy ne csak a kijelölt kifejezés előfordulásait mutassa meg, hanem a CNy szegmentumban is adja meg a kifejezés fordításának közelítő pozícióját. Feltételezésünk szerint lehetséges olyan eljárást kialakítani, amely a fordítók és a fordítással foglalkozó szervezetek számára elérhető korpuszon belül is sok esetben képes a rész-karaktersorozatok CNy megfelelőinek megkeresésére.

5. Fordítástechnológia, terminológia és lexikográfia

5.1. Terminológiai folyamatok a fordításban

A fordítás mindennapi eleme a terminológiahasználat: a fordítók és a fordítással foglalkozó szervezetek szavakat és kifejezéseket gyűjtenek, terminológiai adatbázisokat építenek, és olykor szótárakat is kiadnak.

A terminológiatan irodalma ugyanakkor meglehetős gyéren foglalkozik a terminológia fordítási vonatkozásaival, bár a fordítás kutatói sokszor érintik egyes tárgykörök terminológiai problémáit. Sager (1990) áttételesen másodlagos terminusalkotásról *(secondary term formation)* beszél, amely többek között „történhet [...] másik nyelvi közösség felé irányuló tudásátadás során, terminusok létrehozásával."[30] Arntz (1993) észreveszi, hogy a terminológiai kutatás a fordításban problémát jelenthet: „Hirtelen felmerülő fordítási probléma megoldásához is szükség lehet az adott jelenség részletes tanulmányozására. Az ilyen vizsgálódás gyakran csak a szomszédos fogalmakat említi meg, további részletezés nélkül, így csak az adott tárgykör vagy fogalmi rendszer egy részét kezeli."[31]

Általánosságban azt mondhatjuk, hogy a leíró terminológiatan elsősorban a terminusok (strukturális) nyelvészeti jellemzőivel foglalkozik, alkalmazott tudományként jellemzően azért, hogy valamiféle nyelvészeti modellt találjon a terminusok viselkedésére, amely aztán különféle alkalmazásokban – például a gépi terminuskivonatolásban – felhasználható. A terminológiához kapcsolódó szociolingvisztikai, onomasziológiai vizsgálódások elsősorban előíró jellegűek. Tipikus példa Pavel (1993) írása, amely elviekben a terminológia keletkezésével kapcsolatos neologizmusokat és frazeológiát vizsgálja, ám a folyamatok leíró vizsgálata helyett nyelvoktatási szempontokat ad – egy teljes fejezet foglalkozik az új terminusok helyességének nyelvi feltételeivel. Általában is elmondhatjuk, hogy a terminológiával foglalkozó kutatók jellemzően nómenklátorok vagy meta-nómenklátorok: elsősorban olyan módszereken dolgoznak, amelyekkel lehetséges a fogalmak pontos leírása, megnevezése, illetve amelyekkel ilyenek rendszerét lehet kialakítani.

A terminusalkotás, különösen a másodlagos terminusalkotás azonban igen ritkán történik a teoretikusok által kidolgozott módszerek szerint. A másodlagos terminusalkotás – ahogy a későbbiekből kiderül – elsősorban a fordítás folyamatába ágyazódik. Tény, hogy a legtöbb fordító és fordítással foglalkozó szervezet ezt elvégzi valahogy, anélkül hogy ismerné a teoretikusok módszereit – amelyekről éppenséggel kiderülhet, hogy a fordítás céljainak meg sem felelnek, mert mindig a teljes fogalmi rendszer leírására törekszenek, és emiatt – kü-

lönösen időben – rendkívül erőforrás-igényesek. A terminológiatan adós maradt a terminusalkotás és ezen belül a másodlagos terminusalkotás folyamatainak tanulmányozásával, pedig a terminusok számos nyelvi környezetben az előíró jellegű elmélettől függetlenül is megszületnek. A kérdés szociolingvisztikai és pszicholingvisztikai vizsgálódást egyaránt igényel, az előbbit azért, mert minden terminológia – legalábbis a szociolingvisztika értelmezése szerint – valamely szociolektus, vagyis az adott tárgykörrel foglalkozó vagy vele kapcsolatba kerülő beszélők nyelvhasználatának meghatározó jellemzője.

A következőkben a terminusalkotás folyamatával, azon belül is elsősorban a fordítás során történő terminusalkotással foglalkozom. A fordítástechnológia makrostratégiáinak leírásánál említettem, hogy a terminológiával kapcsolatos stratégiai elemek egyfajta terminológiai munkafolyamatot alkotnak: ennek részletes kifejtése következik most. A fordítástechnológia alapvetően két követelményt állít a terminológiai munkafolyamat elé:

(1) A csoportos fordításban szükség van valamilyen módszerre a konzisztencia biztosításához. A konzisztencia itt azt jelenti, hogy egy adott FNy terminusnak a fordítási munkán belül csak egy CNy megfelelője lehet.

(2) A terminológiakutatás idő- és munkaigényes feladat, ezért a fordítási feladat jobb gazdaságossága végett optimalizálni kell.

Először a terminusalkotás folyamata következik, megkülönböztetve a terminusalkotás három lehetséges színterét: a kutatás-fejlesztési munkát, a szabványosítást és a fordítást.

Másodikként a csoportos fordítási munka terminológiai problémáit írom le két valódi fordítási feladat alapján.

Ezután rendszerszerű áttekintést adok a fordítás terminológiai munkafolyamatáról, annak lehetséges módszereiről és elemi műveleteiről, majd röviden ismertetem az ezt segítő számítógépes eszközöket.

A továbblépés előtt azonban szükség van egy metaterminológiai kitérőre. A ‘terminológia’ szó használata tipikus példa a metaforikus, koordinálatlan terminusalkotásra és -használatra. Ahelyett, hogy részletesen áttekinteném az egyes szerzők szokásait a „terminológia”-használattal kapcsolatban, alább rögzítem, hogy ebben az értekezésben hogyan értelmezem a terminológiával kapcsolatos egyes terminusokat:

Terminológia: adott tárgykör fogalmainak és terminusainak rendszere és e rendszer leírása.

Terminus vagy terminus technicus: adott tárgykör adott fogalmának megnevezése valamely nyelven.

Terminológiatan: a terminológia- és terminusalkotás módszereivel és folyamataival foglalkozó, elméleti és gyakorlati kutatási terület vagy résztudomány.

Terminográfia: a terminológia- és terminusalkotásra, azok közzétételére irányuló tevékenység.

A terminusalkotás folyamata

A tudományban és a technikában a terminológiára azért van szükség, hogy kommunikálni lehessen a szakterülethez kapcsolódó fogalmakról és objektumokról. Amikor új fogalom jön létre, például kutatás-fejlesztési projektek során, új terminus alkotására is szükség van. A terminusalkotás olyan folyamat, amelynek során meghatározott személyek vagy szervezetek valamilyen megnevezést rendelnek adott fogalomhoz vagy objektumhoz.

A terminusalkotást a „mit", „hogyan" és „ki" háromszögével írhatjuk le: személyek vagy szervezetek (ki?) fogalmakat és objektumokat (mit?) látnak el megnevezéssel (hogyan?). A terminusalkotás lehet elsődleges és másodlagos (Sager 1990). Az elsődleges terminusalkotás akkor történik, ha az adott fogalom vagy objektum a terminusalkotók nyelvén korábban nem rendelkezett megnevezéssel, és a terminusalkotók más nyelvi környezetektől függetlenül hozzák létre a terminust. Vitatható, hogy a terminus megváltoztatása az eredeti nyelven elsődleges vagy másodlagos terminusalkotásnak tekinthető-e, Sager a másodlagos terminusalkotáshoz sorolja.

A dolgozat értelmezését megkönnyítendő kizárólag azt a tevékenységet nevezem másodlagos terminusalkotásnak, amelynek során egy létező fogalomnak a forrásnyelven létező megnevezéséhez CNy megfelelőt keresnek.

Az új terminusok létrehozásának három lehetséges színtere van: a kutatás-fejlesztés, a szabványosítás és a fordítás. Felmerülhet egy negyedik is: az oktatás – ugyanis abból a célból is lehet új terminusokat alkotni, hogy lehetséges legyen egy fogalmi rendszert másokkal is megismertetni. Az oktatási anyagok kialakítását ugyanakkor mindig megelőzi valamilyen kutatás-fejlesztési vagy fordítási folyamat.

Terminusalkotás a kutatás-fejlesztésben, az innovációban és a jogalkotásban. E folyamatok során új fogalmak és új objektumok jönnek létre, amelyeket meg kell nevezni, mivel a kutatási eredményeket csak a megnevezés birtokában lehet megosztani a közösséggel.

Az új fogalmak és objektumok megnevezése akkor történik, amikor először kommunikálni kell róluk. Ez kezdetben műhelyen belül történik: a keletkező terminus a műhely tagjai számára érthető, ezért implicit értelmezést igényel, metaforikus és rövid. Ezt nevezhetjük minimális terminusnak. Az új tárgy vagy fogalom azonban csak akkor kerül a fordítás látókörébe, amikor valamely kutatás-fejlesztési projektumról jelentést, konferencia-előadást vagy tanulmányokat írnak, illetve létrehoznak egy jogszabályt vagy szerződést. Ekkor már megjelenik az az igény, hogy a fogalomról vagy tárgyról szóló kommunikáció a külvilág számára is érthető legyen, így az azt megnevező terminusnak szabatos és önmagát magyarázó nyelvi megformálást kell kapnia.

E tekintetben az elsődleges terminusalkotás valójában kétlépcsős folyamat: a műhelyen belüli minimális terminus létrehozása az elsődleges, a publikált terminus létrehozása pedig a másodlagos terminusalkotás. Ilyenformán a szabvá-

nyosítás vagy a fordítás során létrejövő CNy megfelelő kialakítását tulajdonképpen harmadlagos terminusalkotásnak tekinthetjük. Ezt a megnevezést azonban itt nem használom, mert a dolgozatban ezt a terminust a fordítással kapcsolatos terminusalkotás egyik fajtájának tartom fenn – emellett pedig nem foglalkozom azzal a folyamattal, amelynek során a minimális terminusból publikált terminus lesz.

Az új terminusok általában a szakterület elsődleges nyelvén jönnek létre. Más nyelveken ritkán adnak meg ekvivalenseket, hacsak nem írja elő jogszabály, hogy a jelentéseket, tanulmányokat a kutatás-fejlesztést finanszírozó ország nyelvén kell megírni (feltételezve, hogy a szakterület elsődleges nyelve nem azonos a finanszírozó ország hivatalos nyelvével).

A kutatási projektek sokszor nem rendelkeznek módszertannal definíciók írására; nem is születnek mindig formális definíciók. Az új fogalmak és objektumok azonban mindig kapnak implicit definíciót, amelyet a jelentés vagy a cikk szövege fogalmaz meg. Wüster (1979) óta a terminológiaalkotással szemben követelmény, hogy az új terminust explicit és formális definícióval kell alátámasztani, vagyis a terminust a fogalom–név–definíció háromszög határozza meg. Ez jól láthatóan a kutatás-fejlesztés során *sem* történik meg, a terminusok mégis létrejönnek. Sőt: a műhelyen belüli használatra létrehozott, metaforikus minimális terminusok nagyon gyakran a témáról szóló publikált szövegekben is tovább élnek.

A kutatás-fejlesztési projektek során tehát általában nem végeznek szisztematikus terminológiai munkát, a terminusokat intuitív módon rendelik az új fogalmakhoz és objektumokhoz.

Terminusalkotás szabványosítással. A szabványosításnak két célja lehet: a kommunikációs akadályok elhárítása és a nyelvtervezés. A terminológia szabványosítása megkönnyíti a szakmai kommunikációt nemzetközi szinten és a különböző – esetenként különböző szakmai konvenciókkal rendelkező – szervezetek között. Emellett betölt nyelvtervezési szerepet is (Rey 1995:176): adott tárgykörben a nemzeti terminológia, a terminológiai norma kialakítása alkalmassá teszi az adott nyelvet az adott tárgykörrel kapcsolatos szakmai kommunikációra, ezáltal pedig státusztervezési szerepet is betölt.

A szabványosítás feladata az, hogy adott tárgykörben normatív terminológiát alakítson ki meghatározott nyelven vagy nyelveken. Sager szerint a szabványosítás a terminológiaalkotás végső fázisa, „[...] amelynek során a [terminológia] használói közmegegyezésre jutnak arról, hogy adott körülmények között milyen terminust használnak”.[32] A terminológiaalkotás módszertanát az ISO 10241 és az ISO 12615 szabványok is szabályozzák.

A szabványosítási munkát végezheti konzorcium, bizottság vagy munkacsoport, és történhet nemzeti, nemzetközi vagy szervezeti szinten. Az utóbbi jelzi, hogy a terminológia szabványosítása nem kötődik feltétlenül kormányzati szervezetekhez: a nemzetközi cégek maguk is kialakítanak szervezeti szintű

terminológiát, amely a szervezeten belül de facto szabványos, és bár jogi szempontból nem tekinthető szabványosnak, a fordítási folyamatban ugyanolyan, ha nem nagyobb prioritást élvez. A terminológia szabványosításában jelentős szerepet töltöttek és töltenek be nemzetközi szakmai szervezetek. Például az informatika terén a terminológia egységesítését az IFIP-ICC kezdeményezte a 60-as években (vö. IFIP-ICC 1968).

A szabványosítás jól definiált terminusalkotási folyamatokat feltételez. Egy lehetséges folyamat a következő:

1. Definícióalkotás: a terminusok létrehozása előtt definíciók segítségével le kell írni az érintett fogalmakat és objektumokat. A definícióírás valójában a fogalmi rendszer kialakítását, egységesítését szolgálja.
2. Terminusjelöltek meghatározása: ha a definíció által a fogalom vagy objektum már megfoghatóvá vált, a szabványosítási szervezet felméri a lehetséges megnevezések körét.
3. Vita: a jelöltek listája a szabványosító szervezeten belül vita tárgya lesz, amelynek eredménye kötelezően egy és csak egy terminus az adott fogalomra vagy objektumra.
4. Publikálás: a szabványosított terminológia közzététele, általában szótár vagy adatbázis formájában.

A szabványosítás során nem jönnek létre új fogalmak vagy objektumok, vagyis a szabványosítási folyamat mindig olyan entitásokkal foglalkozik, amelyek már rendelkeznek megnevezéssel. A szabványosítási folyamat során a létező megnevezéseket elfogadhatják, módosíthatják, illetve, ha több megnevezés is létezik (amelyeket versengő cégek vagy kutatócsoportok alkottak), közülük egyeseket el is vethetnek, vagyis a meglevő megnevezéseket egységesíthetik.

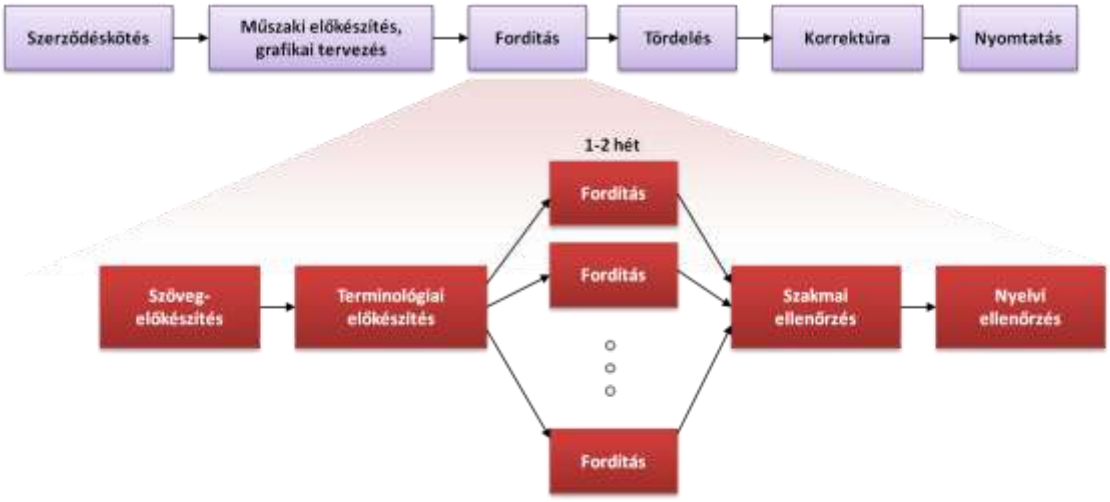
Terminusalkotás a fordításban

A fordításban a terminológiai kutatás a fordítástechnológia kialakulásával kapott nagyobb jelentőséget, pontosabban ennek során merültek fel olyan problémák, amelyek miatt szisztematikus terminológiai munkára lett szükség.

Tekintsünk először két csoportos fordítási feladatot:

1. Könyvfordítás. A feladatot az alábbi adatok jellemzik:

FNy szöveg	(Informatikai) szakkönyv
Terjedelem	500 oldal, kb. 180 000 FNy szó
Tevékenység	fordítás; minőségellenőrzés; korrektúra; tördelés
Elvárt kimenet	nyomdakész CNy szöveg
Határidő	a kezdéstől számítva 4 naptári hét



5.1. ábra: Könyvfordítás lehetséges munkafolyamata

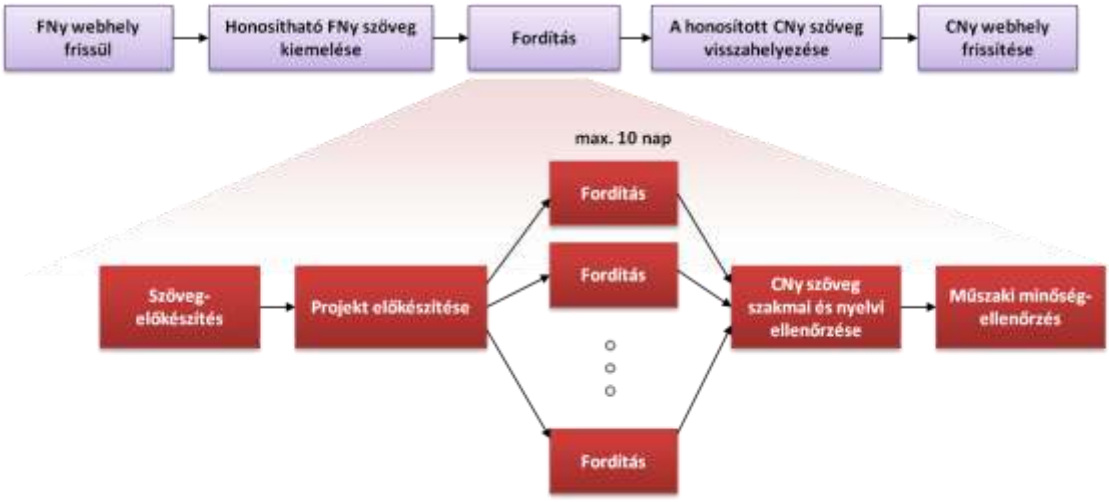
Az 5.1. ábrán a könyvkiadási folyamat lineáris munkafolyamatának áttekintése látható. A szűk határidő miatt a könyvkiadó párhuzamosítja a fordítást: a FNy szöveget 4-7 részre osztja, így a fordítás 1-2 naptári hét alatt elvégezhető.

A kiadott könyvvel szemben rendkívül szigorú minőségi követelmények állnak fenn, ezért a folyamat minden olyan minőségbiztosítási műveletet tartalmaz, amelyet egy kiadó el szokott végezni: szakmai ellenőrzés, nyelvi ellenőrzés és korrektúra. Az egységes terminológiát a kiadó azzal igyekszik biztosítani, hogy a fordítás megkezdése előtt terminológiai adatbázist épít a FNy szöveg tartalmából.

Az 5.1. ábrán nem látszik, hogy a fordítás és a minőségellenőrzés fázisa átfedhetik egymást. A FNy szöveg felosztása a könyv fejezetstruktúrája alapján történik. A 2 hétnyi munkaidőre minden fordító legalább 2 részt kap, így már az első hét végén lehetőség van CNy szöveg átadására: ezek esetében a minőségellenőrzés azonnal elkezdődhet.

2. Webhely-honosítás. A feladat egy összetett webhely lefordítása, amelynek során a webhely műszaki megvalósítását meg kell őrizni. A feladat főbb jellemzői:

FNy szöveg	Szakmai webhely
Terjedelem	kb. 120 000 FNy szó
Tevékenység	fordítás és minőségellenőrzés
Elvárt kimenet	CNy webhely tartalma
Határidő	a kezdéstől számítva 12 naptári nap
Erőforrás	a webhely korábbi tartalma, fordítómemóriában



5.2. ábra: Webhely-honosítás lehetséges munkafolyamata

Az 5.2. ábrán látható lineáris munkafolyamat lényegében megegyezik a könyvkiadási projektével. Két olyan feladat van, amely nem kapcsolódik a fordításhoz vagy a terminológiához:

- a FNy szöveg műszaki előkészítése a fordítók számára és
- a lefordított CNy szöveg műszaki minőségellenőrzése: ez arra biztosíték, hogy a webhely tartalmába foglalt kódelemek eredeti formájukban a CNy szöveg megfelelő pozícióira kerülnek.

Megfigyelhetjük, hogy a második fordítási feladat a tartalom jellege miatt közelebb áll a technikához, elvégzéséhez emiatt is több gépi eszközre van szükség.

A fordítás terminusalkotás munkafolyamata

A fordítást mindkét fordítási feladat esetén párhuzamosítani kell, másképp nem végezhetők el. A fordítást több fordító végzi, egymástól függetlenül. Ezzel a módszerrel a legnagyobb minőségi kockázat – amely a CNy szöveg olvashatóságát is veszélyezteti – a terminológiahasználat és a stílus konzisztenciájának hiánya. Mindkét munkafolyamat magában foglal terminológiai előkészítést, de mi történik, ha néhány tucat terminus nem kerül az előkészített szószedetbe? Már egyetlen terminus félreértelmezése is negatívan befolyásolhatja a CNy szöveg érthetőségét.

A konzisztencia minőségi követelménye (Kis Á.–Kis B. 2003) vajon azt jelenti-e, hogy a CNy szövegben a terminusoktól elvárjuk a szinonimamentességet? Ez a FNy (vagyis bármely nyelven megírt eredeti) szakmai szövegeket éppen nem jellemzi (Sager 1990:59, 214) – így a lefordítandó FNy szöveget sem feltétlenül. A forrásnyelvi szövegben ráadásul lehetnek szabályosságok, amelyek alapján különböző környezetekben a FNy terminusnak (pontosabban a fogalom megnevezésének) más és más variánsa fordul elő.

A konzisztencia ezért tulajdonképpen nem a homogén, hanem a szabályozott terminológiahasználatot jelenti: a követelmény nem az, hogy adott fogalomra

mindig ugyanazt a terminusvariánst kell használni, hanem az, hogy a koordinált használati szabályoktól nem szabad eltérni. A problémát elsősorban nem az jelenti, hogy ugyanazt a FNy terminust két különböző variánssal fordítják (ez legfeljebb zavaró, ha nem követ szabályosságot), hanem az a kockázat, hogy a fordítók nem feltétlenül szakértői a fordítandó szöveg szakterületének, ezért a koordinálatlan fordítás félreértelmezést eredményezhet. A szinonimamentesség új fogalmak bevezetése esetén követelmény: ha az olvasónak más forrásból nincs lehetősége azonosítani a CNy vagy a FNy terminus által jelölt fogalmat, akkor a szövegben (akár FNy, akár CNy szövegről van szó) a variánsok használatát explicitálni kell.

A fordítás során végzett a terminusalkotás feladatát kétféleképpen is meg lehet fogalmazni:

- CNy megnevezések hozzárendelés meglevő fogalmakhoz;
- a FNy terminusokat CNy megfelelőikkel kell helyettesíteni.

Bár a két megközelítés nem teljesen ekvivalens, csak abban a nyelvfüggő stratégiában különböznek, amely az egyes terminusok megalkotását szabályozza. Mindkettő eredményezhet helyes és egységes terminológiahasználatot.

A „helyes" meglehetősen homályos kifejezés – épp emiatt nem tekinthető terminusnak sem –, ha konkrétabban akarunk fogalmazni, a CNy terminustól azt kell elvárnunk, hogy megfeleljen

- a hatályos szabványoknak,
- a szakterület de facto szabványainak (amelyet a főbb piaci szereplők vagy konzorciumaik határoztak meg),
- a FNy szöveget kibocsátó vagy felhasználó szervezet terminológiájának,
- az ügyfél (a fordítás megrendelője) által előírt terminológiának.

A fordításban ezért a konzisztencia mellett a szabályoknak való megfelelés is fontos követelmény. Éppen emiatt Lengyel (2005) külön fordítási terminológiáról beszél, amelybe a FNy szöveg minden olyan lexémája beleértendő, amelyet egységesen, illetve meghatározott szabályok szerint kell fordítani.

A fordításban a terminológiai munka másodlagos és harmadlagos terminusalkotást jelent. A harmadlagos terminusalkotás alatt a fordítótól vagy a fordítást végző szervezettől függetlenül már létező CNy terminus megkeresését értem. Az elsődleges terminusalkotást kizárhatjuk, mert a FNy terminus nyilvánvalóan léteznek, vagyis a szövegben megjelölt fogalmak és objektumok már kaptak megnevezést a forrásnyelven. Azonban a gyorsan változó szakterületek esetén valószínű, hogy számos FNy terminusnak nem létezik vagy nincs elterjedt, illetve szabványos CNy megfelelője.

A fordítási projektekben két terminológiai feladat van:

(1) A FNy terminusok megkeresése a szövegben;

(2) CNy megfelelők keresése a FNy szövegben talált terminusokhoz.

A FNy terminusok megkeresése a szövegben. A fordítás során minden FNy terminust észre kell venni a szövegben. A gyakorlatban ez azt jelenti, hogy minden szót és kifejezést fel kell ismerni, amelyet egységesen vagy meghatározott szabályok szerint kell fordítani. A minimális fordítás (Heltai 1999) megközelítése alapján viszont a terminológiai glosszáriumnak azokat és csak azokat a kifejezéseket (lexémákat) kell tartalmaznia, amelyek egységes, illetve szabályoknak megfelelő fordítása elengedhetetlen a CNy szöveg megértéséhez.

Ha a szövegben kijelöljük az összes terminust, megkapjuk a szöveg terminológiai vázát. Így a szöveg tulajdonképpen két részre osztható: a terminológiai vázra és a diskurzusstruktúrára (Kis Á. 2002, Kis B. 2005a). Ideális esetben a fordítás során a teljes terminológiai vázat le kell képezni a célnyelvre, vagyis az összes FNy terminust meg kell találni. Valószínű, hogy a meg nem talált FNy terminusok fordítása egységes sem lesz, és a szabályoknak sem felel majd meg.

CNy megfelelők keresése a FNy szövegben talált terminusokhoz. Erre számos stratégia létezik, ezek azonban egyformán munkaigényesek. Saját tapasztalataink szerint a terminológiai kutatás a fordításra fordított idő 40-60%-át is igényelheti.

Mivel egyfelől a terminológiai munka a fordítási munka jelentős részét kitöltheti, másfelől pedig a szakmai fordítás minősége nagyrészt a terminológiahasználat egységességén és megfelelőségén múlik, létfontosságú, hogy a fordítási projekt szisztematikus és jól definiált terminológiai stratégiával rendelkezzen.

A terminológiaalkotás stratégiája a fordításban

A terminológiai munka stratégiája két szinten definiálható:

a) a projekt szintjén: ez a terminológiai munkafolyamat (tulajdonképpen „makrostratégia”);
b) a terminus szintjén: ez a terminusalkotási stratégia („mikrostratégia”).

A fordítástechnológiát az előbbi érdekli – az utóbbi széles körű nyelvészeti vizsgálódás tárgya, amelyhez ez az értekezés nem tudna sokat hozzátenni (lásd pl. Sager 1990:60-89, Rey 1993:105-112, Zauberga 2005).

Feltételezzük, hogy a fordítási munkát fordítócsapat végzi, amelyben fordítók, egymástól függetlenül, ugyanazon szöveg vagy szövegegyüttes különböző részeit fordítják. Nyilvánvaló, hogy a terminológiahasználat egységességének és megfelelésének biztosítása többletmunkával jár, és többletidőt is igényel – a párhuzamosított, de koordinálatlan fordításhoz képest. Ebben a felállásban a csoport több különböző stratégiát követhet. Mivel a fordítás – rendeltetését tekintve – gazdasági tevékenység, a stratégia kiválasztása az idő–költség–minőség háromszög mentén kialakított kompromisszum eredménye (Lewis 2005). „A kompromisszum háromszöge, amelynek csúcsai az ütemezést, a költséget és a minőséget képviselik, általános vezetési-szervezési alapelv.” (McConnell 1996:126)[33]

A stratégia kialakításához a három tényező közül egyet rögzítünk: a minőséget. Feltételezzük, hogy a csoport mindig a maximális minőség elérésére törekszik. Ez ellentmondásnak tűnhet, hiszen a fordítástechnológiát épp az hívta életre, hogy a költség és az idő vált (rendkívül szűk keretek között) rögzített paraméterré. A stratégia kialakításához azonban épp azt a kérdést kell megválaszolni, hogy a maximális minőségre törekvés mellett miként lehet a szűkös idő- és költségkereten belül maradva elvégezni a munkát.

A választható stratégiák elhelyezhetők egy időtengelyen. Az adott stratégia időtengelyen felvett pozíciója azt jellemzi, hogy a terminológiai munka legnagyobb részét a fordításhoz képest mikor hajtják végre:

5.3. ábra: A fordításban alkalmazott terminológiai stratégia skálája

Az időtengely két végpontja a terminológiai stratégia két szélsőségét jelzi. A következőkben ezeket foglalom össze a lehetséges köztes stratégiával együtt.

Teljes előkészítés. Ennek során a lehető legteljesebb terminológiai szószedetet előállítják a fordítás megkezdése előtt. Ekkor végeznek el minden terminológiai kutatást, és ekkor teremtik meg a műszaki feltételeket ahhoz, hogy a fordítók ne térhessenek el a glosszáriumban előírt CNy terminusok használatától. Az ideális esetben ellenőrizni kell, hogy a FNy terminusok halmaza teljes-e, a CNy terminusokat pedig ellenőrizni és korrektúrázni kell. Ez a folyamat rendkívül időigényes, mert a FNy szöveget végig kell olvasni, ráadásul valószínűleg többször. A többlet-időigény hatása azért is nagy, mert nem kezdődhet meg a fordítás, amíg a teljes glosszárium nincs befejezve.

Az előkészítési fázist ugyanakkor lehet gépi eszközökkel és jól definiált módszertannal támogatni (lásd a 18. ábrát!). Az előbbivel az 5.2. fejezet bővebben is foglalkozik.

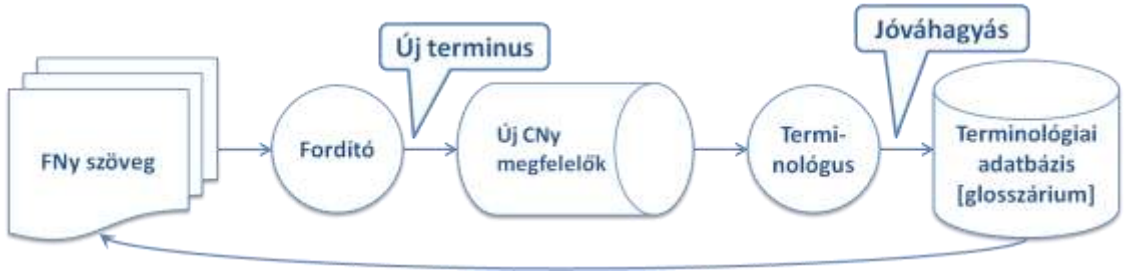
5.4. ábra: A teljes előkészítés stratégiájának lineáris munkafolyamata

Teljes utólagos ellenőrzés vagy teljes lektorálás. E stratégia esetén nem készül glosszárium a fordítás előtt. A fordítók egymástól függetlenül végeznek terminológiai kutatást. A projektben olyan lektor dolgozik, aki a fordítás minden aspektusát ellenőrzi és javítja, beleértve a terminológiahasználat egységességét és megfelelését. Ez a megközelítés rendkívül költséges, mert ekkor kell a legtöbbet módosítani a fordítótól származó CNy szövegen. Tapasz-

talatunk szerint a lektoráláshoz átlagosan harmadannyi idő szükséges, mint a FNy szöveg lefordításához; ebben az esetben azonban lényegesen több időre is szükség lehet.

Felügyelt együttműködés. A valós fordítási munkafolyamat a terminológiát illetően általában a két szélsőség közé esik. A lineáris fordítási munkafolyamatban előkészítési és ellenőrzési fázis is van. A legalacsonyabb költséget azonban a felügyelt együttműködés esetén lehet elérni.

A nagy, de szűk határidejű fordítási munkát általában párhuzamosítjuk: több fordító dolgozik egyszerre. A felügyelt együttműködés a terminológiai munkafolyamatban valósítja meg ugyanezt: a terminológiai kutatómunkát is párhuzamosítja. A 5.5. ábrán ennek egyszerűsített vázlata látható.



5.5. ábra: A felügyelt együttműködés vázlata

A felügyelt együttműködés során lehetőség van glosszárium előkészítésére a fordítás előtt. A terminológiai kutatás nagy része azonban a fordítókra marad – ez tulajdonképpen hasonlít a teljes lektorálás modelljéhez. A legfontosabb különbség az, hogy a fordítók többé nem függetlenül dolgoznak: sem egymástól, sem a terminológustól. Ez a modell terminológiai adatbázist (a glosszáriumot) dinamikusan kezelt közösségi erőforrásnak tekinti. Amikor egy fordító befejezi a kutatást valamelyik terminussal kapcsolatban, felvesz egy új szócikket a közös terminológiai adatbázisba. Az adatbázisnak van lektora – a terminológus –, aki rendszeresen áttekinti az újonnan felvett szócikkeket, és szükség esetén javítja őket.

Ez a stratégia a következőképpen biztosítja a terminológiahasználat egységességét és szabályoknak való megfelelését:

- Az új szócikkek azonnal láthatóvá válnak a csoport többi tagja számára. Ha a csoport tagjai olyan fordítási környezetben dolgoznak, amely kiemeli a szövegből az adatbázisban szereplő FNy terminusokat, a fordítók adott FNy terminus minden további előfordulására automatikusan megkapják az első fordító által kidolgozott CNy megfelelőt. Ezzel biztosítottuk a terminológiahasználat egységességét.
- Az új szócikkeket a terminológus minden esetben ellenőrzi és javítja, vagyis azok mégsem azonnal, hanem egy kis késleltetéssel válnak láthatóvá a fordítók számára: akkor, amikor a terminológus jóváhagyta vagy kijavított az el-

ső fordító által felvett CNy megfelelőt. Ez biztosítja, hogy a terminológiahasználat megfeleljen a szabályoknak.

A „felügyelt együttműködés" terminus feltételezi, hogy a csoporton belül létezik kommunikáció – amely nélkül nem lehetséges az „együttműködés". A munkafolyamat vázlata (lásd az 5.5. ábrát!) ezt nem mutatja, csak arra a korlátozott kommunikációra utal, amelynek során a terminusjelöltek „felfelé" (a terminológiai adatbázis felé), a visszacsatolás elemei (a javított terminusok) pedig „lefelé" (a CNy szöveg felé) haladnak.

A folyamat ugyanakkor azt is feltételezi, hogy a projekt résztvevőit valamiféle más kommunikációs infrastruktúra is összeköti. Így olyan általános kommunikációra is lehetőség van, amelyben minden fordító részt vesz, a terminológus pedig moderátorként viselkedik. Ez a következő előnyökkel jár:

- A terminológiai kutatás okozta terhelés jobban eloszlik a csoport tagjai között. Egyetlen fordító és a terminológus sincs magára hagyva a helyes CNy megfelelők megkeresésében.
- A terminológiai kutatással töltött idő beépül a fordítási időbe, és az is eloszlik a fordítók között. Ha azzal számolunk, hogy a csoportos fordítási munkák terjedelme legalább 200 000 leütés (40 ezer szó), ahol a FNy szöveg kb. 1000 fordítási értelemben vett terminust tartalmaz, akkor valószínű, hogy a terminológiai munka 3-7 fordító között egyenletesen oszlik el. Ez azt jelenti, hogy a felügyelt együttműködés stratégiája adja hozzá a lehető legrövidebb időt a fordítás idejéhez.
- Javul a fordítás során kialakított CNy megfelelők minősége, mert érvényesül a „több szem többet lát" effektus: az általános kommunikációs infrastruktúra (e-mail, vitafórum, csevegőszoba) lehetővé teszi, hogy a csoport egyetlen CNy megfelelőt is kollektív módon alakítson ki. Az ilyen jellegű terminológiai vitákra példa a Microsoft honosítási műhelyének gyakorlata, amely új, nagyobb rendszerek honosításakor a felhasználói-szakértői közösség terminológia iránt elkötelezett tagjait bevonja az új terminológia kialakításába.[34]

A terminológiakezelés eszközei a fordításban

A technológiából az eddigiekben a folyamatokat írtam le. A leírás azonban feltételezi bizonyos eszközök használatát, amelyek nélkül például a felügyelt együttműködés stratégiáját nem lehet alkalmazni.

A fordításon belüli terminológiai kutatás fő problémája az, hogy a helyes CNy megfelelők megkeresése munkaigényes. A fordítók, fordítással foglalkozó szervezetek sok időt és pénzt áldoznak arra, hogy megkeressenek olyan CNy terminusokat, amelyek már léteznek, de nem triviálisan elérhetők a fordító vagy a fordítást végző szervezet számára.

Ezért szükséges, hogy a fordítástechnológia eszközei és folyamatai megkönnyítsék

(1) az előkészített terminológiához való hozzáférést,

(2) a terminológiai egységességgel és megfeleléssel kapcsolatos minőségbiztosítást,

(3) a fordítócsoport tagjai közötti kommunikációt és

(4) a terminológiai erőforrások szétosztását vagy közzétételét.

A következőkben röviden összefoglalom, hogy a jelenlegi gépi eszközök hogyan felelnek meg a fenti követelményeknek.

(1) Hozzáférés az előkészített terminológiához. Ennek biztosítása a fordítási környezet, vagyis a fordítás megírásához használt program feladata. A probléma az, hogy egy 500 oldalas könyvön végzett terminuskivonatolás eredménye kb. 2500 ellenőrzött terminuspár (szócikk). Ennek megtanulását nem lehet elvárni a fordítótól: segítség nélkül a fordító nem tudja megállapítani, hogy az aktuális FNy szegmentumban a terminológiai adatbázis mely elemei fordulnak elő.

A fordítási környezet feladata ekkor az, hogy folyamatosan figyelje a terminológiai adatbázist, és a FNy szövegben eltérő színnel emelje ki az adatbázisban megtalált terminusokat. Többletszolgáltatásként lehetőséget adhat arra is, hogy az adatbázisban tárolt CNy megfelelőt a fordító egy lépésben a szövegbe illessze. Ezzel a fordítót a következő, relatíve időigényes lépésektől kíméli meg:

- a FNy szegmentumban előforduló potenciális terminusok kijelölése;
- átlépés más programba, a terminusjelölt megkeresése;
- a megtalált CNy megfelelő beillesztése.

Az értekezés írása idején már a legtöbb fordítási környezet végez terminuskiemelést, azonban megfigyelésem szerint a legtöbb fordítással foglalkozó szervezet nem alkalmaz terminológiai adatbázist.

(2) Minőségbiztosítás. Léteznek konzisztencia-ellenőrző programok (ezzel a 3.2. fejezetben foglalkoztam), amelyek azonban sok fordítási környezetben nem érhetők el; a meglevő programok szolgáltatásköre pedig egyelőre bővítésre szorul.

Az ilyen eszközök annyiban gyorsítják a minőségbiztosítási munkát, hogy – adott esetben még a fordítás közben – megjelölik azokat a szegmentumokat a CNy szövegben, ahová nem került be a FNy szegmentumban előforduló terminus egyik előírt CNy megfelelője sem. Megbízható eszköz esetén ezzel megkímélhetik a lektort a szöveg terminológiai szempontú átolvasásától, ha kevés az idő: ilyenkor elég lehet a jelzett szegmentumok áttekintése. [Ismét utalok az idő–költség–minőség háromszögre, illetve a Heltai (1999) által bevezetett minimálisfordítás-koncepcióra.]

(3) Kommunikáció és együttműködés a csoporttagok között. Az általános kommunikációs infrastruktúra (e-mail, vitafórum, csevegés) mellett hasznos eszköz a moderált terminológiaépítés, illetve a moderált terminológiai adatbázis. Erre alapvető szükség van a felügyelt együttműködés stratégiájához.

Ugyanitt érdekes lehet a formális kérdéskezelés. Ez hasonlít a moderált terminológiához: a különbség az, hogy a fordító nem új terminuspárt javasol, hanem megoldatlan terminológiai kérdést tesz fel, amelyre jellemzően a terminológus válaszol. Ilyen rendszert egyes fordítóirodák üzemeltetnek, azonban nyilvánosan elérhető kérdéskezelő *(query management)* rendszert még nem láttam.

A kérdéskezelő rendszer kétféle megközelítést alkalmazhat: lehet a terminológiai adatbázishoz hasonló, ahová a fordító úgy vehet fel FNy terminust, hogy a CNy terminus helyett kérdést és előfordulási példát tölt fel. A másik megközelítés a szoftverfejlesztés minőségbiztosításában alkalmazott hibakövető (issue management) rendszerekéhez hasonló: itt a terminológiai problémát eseménynek tekintik, és akként dolgozzák fel. Az utóbbi hátránya, hogy viszonylag nehéz megkeresni az adott FNy terminusra vonatkozó kérdést, mert az adatbázis általában a kérdések ideje és státusza, nem pedig a kérdéses FNy terminus szerint van rendezve.

A kérdéskezelő rendszer a terminológiai kutatást könnyíti meg: a kutatási feladat könnyen átadható annak a személynek, aki rendelkezhet a CNy terminus meghatározásához szükséges előismeretekkel, így a terminológiai kutatás időigénye összességében csökken.

(4) A terminológia szétosztása vagy közzététele. A műszaki-tudományos terminológia jelentős része nem érhető el szótárban, a szakma- és szervezetspecifikus terminológiai gyűjtések pedig vagy nem nyilvánosak, vagy nem egységes felületen érhetők el. Ezért számos – tulajdonképpen szabványosítási – törekvés született

- a hálózaton keresztül elérhető terminológiai erőforrásokhoz való hozzáférés egységesítésére és fordítási környezetekbe integrálására, illetve
- a terminológiai erőforrások egyesítésére.

Az előbbi törekvés – amennyiben az erre irányuló, az értekezés befejezése idején kezdődő fejlesztés sikerrel jár – azt célozza, hogy az interneten szétszórt heterogén erőforrások ugyanúgy vagy majdnem ugyanúgy legyenek elérhetők a fordítási környezetekből, mint a belső terminológiai adatbázisok.

Az utóbbira pedig sikeres példa a nemzetközi együttműködésben végrehajtott EuroTermBank projektum. A terminológiai erőforrások egyesítése azt jelenti, hogy heterogén terminológiai adatbázisokat (amelyek esetleg nem is egy rendszerben vannak tárolva) úgy teszünk hozzáférhetővé, hogy a felhasználó számára a rendszer egy nagy terminológiai adatbázisnak látszik. Az EuroTermBank rendszer[35] (Rirdance-Vasiljevs 2006) ennek során még szócikk-egyesítést is végez: ha több különböző forrásadatbázisban, egy adott nyelven szerepel ugyanaz a terminus, akkor a közös terminus – és még néhány járulékos szabály – alapján, a keresés után egyesíti a szócikkeket. Ez dinamikusan zajlik: ha új adatbázist adunk a rendszerhez, az egyesített szócikk legközelebb már esetleg másképp jelenik meg.

5.2. Terminuskivonatolás

A terminuskivonatolással kiterjedt nemzetközi irodalom foglalkozik (pl. Jacquemin 2001), és számos alkalmazása is született. A következőkben egy fejlesztési projektről és kísérletről számolok be, amely arra irányult, hogy nagy fordítási feladatokhoz használható terminuskivonatoló eszköz jöjjön létre, amelynek segítségével

- leképezhető a FNy szöveg teljes terminológiai váza, és
- jelentősen rövidíthető a terminológiai előkészítésre fordított idő.

Az utóbbi kritérium konkrétabban úgy fogalmazható meg, hogy a szövegspecifikus terminológiaalkotáshoz szükséges időnek lényegesen rövidebbnek kell lennie, mint amennyi idő alatt a FNy szöveget egy ember figyelmesen végig tudja olvasni.

Fontos megjegyezni, hogy a létrehozott terminuskivonatolási technológia elsőként támogatja a magyar nyelvet; a rugalmas nyelvészeti keretrendszer pedig megkönnyíti újabb nyelvek hozzáadását.

Az itt következő leírás az erre irányuló korpusznyelvészeti kutatást írja le, és megmutatja, hogy a létrehozott eszköz mindkét fenti követelménynek megfelel. A leírás elején ismét foglalkozom a terminológia definíciós problémáival, de ezúttal a gépi modellezés szemszögéből.[36]

A terminológia modellezése

Definíciós nehézségek. A fordítás szempontjából a terminológia definiálása azért nehéz, mert a fordítás során már nem szabályozható a FNy szakszöveg létrehozása, s így annak terminológiahasználata sem. Ezért teszünk különbséget elsődleges és másodlagos terminusalkotás között.

Problémát jelent a FNy szövegek interdiszciplináris jellege is. Az ideális esetben a terminológia minden tárgykörben szabványosítva van. Jelentse ez itt a fogalmak és tárgyak definiálását, illetve normatív – adott esetben többnyelvű – megnevezését. A tárgykört azonban nehéz körülhatárolni, mert a terminológiát használó diszciplínák nem szigetelődnek el egymástól. A tárgykörök összefüggései bonyolult ontológiai hierarchiát alkotnak (vagy inkább hierarchiák halmazát). Adott szakmai szöveg pedig – különösen az alkalmazott tudományokban – mindig interdiszciplináris lesz, vagyis nem tudunk felmutatni egy jól meghatározott tárgykört, amelybe a szöveget besorolhatnánk. Ebben az esetben pedig a FNy szöveg az összes érintett tárgykör terminológiáját használja, s mivel a szabványosítás során az egymástól eredetileg – ontológiai szempontból – távol eső diszciplínák terminológiáját egymástól függetlenül alakítják ki, a terminológiák egyesítésekor olyan ütközések fordulhatnak elő, amelyek többértelműségeket, inkonzisztenciákat okozhatnak (vö. Sager 1990:58).

A szakmai fordítás gyakorlata azonban azt is bebizonyította, hogy terminológia nemcsak szabványosítással, hanem intuitívan is keletkezik. Ez azt jelenti,

hogy sok terminus nem explicit definiálás útján jelenik meg először a szakmai kommunikációban, hanem szakszövegekbe foglalva, implicit definiálás útján is. Ahhoz viszont, hogy az ilyen terminusok valóban terminológiává váljanak, szükséges, hogy mind az eredeti szöveg(ek) szerzője (szerzői), mind pedig a szakmai kommunikációban részt vevő partnerek – formailag és szemantikailag – konzisztensen kezdjék használni.

Megjegyezzük, hogy a fenti állítások tapasztalati tényeken, nem pedig szigorú tudományos felmérések eredményein alapulnak. Ez utóbbi lehetne a terminológiahasználat szociolingvisztikájának kutatása, amely azonban nem tárgya ennek az írásnak. Fordítói körökben mindazonáltal mindennapi tapasztalat, hogy a forrásszövegekben százával jelennek meg olyan szavak és kifejezések, amelyeket környezetük alapján egyértelműen terminusnak kell minősíteni, mégsem találhatók meg egyetlen, a fordító számára hozzáférhető szabványos vagy legalábbis autentikus szótárban sem.

Ha a szakmai kommunikációban alapkövetelmény a konzisztens terminológiahasználat, akkor ez érvényes a szakmai fordításra is. Azonban a fordítás nem forrása vagy befogadója, hanem „csupán" csatornája a szakmai kommunikációnak. A fordító maga többnyire nem kompetens szakmai kommunikátor, hanem mediátor, akinek sem eszköze, sem ismerete, sem erőforrása nincs elegendő ahhoz, hogy az adott forrásszövegben megjelenő diszciplínák teljes fogalmi hálóját – a benne levő terminológia teljes szemantikai tartalmát – birtokolja. [Ez egyébként a fordítás kutatásának alapvető problémája: milyen „mélységig" kell, illetve lehet megérteni a forrásszöveget a sikeres fordításhoz? (vö. Komisszarov 1990 az ekvivalencia szintjeiről)]

A terminológia definiálása a fordítás szempontjából. A fordító mint mediátor szempontjából tehát érdektelenek a terminológia keletkezésének körülményei, mert a terminusokat szakszövegbe ágyazva, készen kapja. Feladata az, hogy ezeket felismerje, és konzisztens módon fordítsa a célnyelvre.

Ezért a fordítás szempontjából a terminusnak egyetlen ismérve van: konzisztensen (egységesen és meghatározott szabályoknak megfelelve) kell fordítani. A terminológia most nem szótárakban, definíciókban, adatbázisokban létező elvont fogalom: a terminus egyértelműen szövegnyelvészeti jelenség, amelyről korábban már leírtuk, hogy alapvetően két attribútuma által definiálható (Kis Á. et al. 2004):

- a terminológiai helyzet: ez az a szövegnyelvészeti jelenség vagy jelenségrendszer, amely alapján a szövegben felismerhetők azok a szavak vagy kifejezések, amelyeket terminusként kell kezelni;
- a terminológiai szerep: a felismerhetően terminológiai helyzetben álló, tehát terminusként kezelendő szó vagy kifejezés egyértelműsége sérülhet a szöveg interdiszciplináris vagy multidiszciplináris jellege miatt. Így sok esetben feladat a megfelelő terminológiai szerep kiválasztása. Ez egy-

mástól távol álló és egymáshoz rendkívül közel eső tárgykörökben is előfordulhat (pl. a morfológia szót számos különböző tudományban használják, ugyanakkor pedig pl. a *'directory'* szó (angolul) az informatika két, egymáshoz közel eső területén mást jelent).

Létezik a terminológiával kapcsolatban egy olyan előíró jellegű közmegegyezés, amely szerint a terminusok monoszémiája és egyalakúsága (homonimamentessége) szükséges a szakmai szöveg megértéséhez. (Kis Á, 2003:49) Ez így általában nem valósul meg, pontosabban mindig csak relatív: a kontextus egy adott lexémát terminológiai helyzetbe hoz, amely a környezet hatására felveszi a terminológiai szerepet, vagyis azt a jelentéskört, amely a terminológiában meghatározott fogalomra fókuszál (ismét Kis Á, 2003:49), így lokálisan a monoszémia és az egyalakúság is fennáll. Sager (1990) leíró megközelítése már a lokális egyalakúságot említi: „Az onomasziológiai megközelítésben és a CNy megfelelők keresésekor a különböző alakok közül kontextuális minták alapján kell választani, ami használati szabályokra fordítható le." (Sager 1990:59)[37]

A terminológia modellje a számítógép szempontjából

A terminuskivonatolás részben alkalmas a fordítás terminológiai előkészítésének automatizálására. Ha azonban automatizálásról beszélünk, halmozottan érvényesek lesznek azok a szempontok, amelyek a nyelvtechnológiai fejlesztések során mindig előkerülnek. A nyelvtechnológia ugyanis nem tanítja meg a számítógépet a természetes nyelv megértésére és feldolgozására: a tudomány jelenlegi állása szerint – sem az informatikában, sem a kognitív tudományokban – nincsenek meg az ismereti és tárgyi feltételei annak, hogy az emberi nyelvprodukcióhoz, illetve -befogadáshoz szükséges kognitív hátteret át tudjuk adni a számítógépnek. A nyelvtechnológiai alkalmazások mindig csak utánozzák az ember egyes, parciális nyelvi funkcióit, és igyekeznek olyan módszereket alkalmazni, amelyek nem igénylik a feldolgozandó szövegek értelmezését (vagyis szemantikájának feldolgozását).

A nyelvtechnológiai kutatók rendszerint olyan – lexikális, morfoszintaktikai, szintaktikai – felszíni jelenségeket keresnek a szövegekben, amelyek egyszerű, kis erőforrás-igényű algoritmusokkal felismerhetők. Az egyszerűségre törekvés azért is fontos, mert a nyelvtechnológiai modulok gyakran csak kiegészítő, kiszolgáló funkciót töltenek be komplex informatikai rendszerekben, s így nem engedhetik meg maguknak, hogy lekössék a rendszert futtató számítógép kapacitásának jelentős részét. Valójában az általuk okozott terhelés rendszerint még a kapacitás 10%-át sem érheti el, jellemzően elvárás, hogy a nyelvi alrendszereknek a teljesítmény szempontjából észrevehetetleneknek kell maradniuk.

Ha pedig a nyelvi jelenségeket a számítógépen rendszerint felszíni jegyekkel modellezzük, semmiféle kognitív modellt nem alkalmazhatunk. Ezért az emberi nyelvi funkciók számítógép általi „utánzása" is csak közelítőleg lehetséges, vagyis az emberi funkció felszíni viselkedését közelítjük a számítógéppel.

Az egyszerűsített algoritmusokat a korpusznyelvészet eszközeivel dolgozzuk ki, és ugyanezen eszközökkel igazoljuk megfelelőségüket. Az itt következő módszerek mind ennek a szemléletnek felelnek meg, s bár kognitív szempontból már az első ránézésre inadekvátak, gazdasági jelentőségük nagy, mivel az ember munkáját mégis lényegesen hatékonyabbá teszik.

A terminuskivonatolás módszereinek áttekintése

A gépi terminuskivonatolás technikailag azt jelenti, hogy a forrásszöveget bemenetként adjuk egy programnak – vagy több, láncba fűzött programnak –, az eredmény pedig olyan szavak és kifejezések listája, amelyeket az algoritmus lehetséges terminusnak minősített. A lista esetleg kiegészülhet statisztikai, kontextus- vagy morfoszintaktikai adatokkal is.

Fontos, hogy az önálló szövegelemzést, illetve -kivonatolást végző programok konfidenciaszintje igen alacsony. Ez azt jelenti, hogy a programok kimenetét emberi utófeldolgozás nélkül nem lehet megbízhatónak tekinteni. Kutatásunk nem arra irányul, hogy az emberi utófeldolgozást eliminálja a folyamatból – ez a legföljebb megközelíthető ideális eset –, hanem arra, hogy minimalizálja az ezzel járó munkát.

A fentiek miatt a programok által kimenetül adott listákban nem terminusok, hanem úgynevezett terminusjelöltek jelennek meg, amelyek terminus voltát igazolni vagy cáfolni kell (Jacquemin 2001; Castellví et al. 2001).

Feltételeztük, hogy kockázatos egy algoritmusra támaszkodni, tudván, hogy minden felszíni közelítő módszer jelentősen túl-, illetve alulgenerál. Ezért úgy döntöttünk, hogy több módszer kombinációját alkalmazzuk. Ezek között két statisztikai és két szabályalapú algoritmus található. A módszerek kombinálásával azt is szeretnénk elérni, hogy a különböző algoritmusok végezzék el helyettünk a terminusjelöltek egy részének megerősítését vagy elvetését.

Szabályalapú módszerek. Szabályalapú eljárásainkhoz a forrásnyelvek morfológiai elemző programja, korlátozott méretű alapszótára, illetve sekély (lokális) mondatelemző programja szükséges, ezért kezdetben ezeket csak a magyar és az angol forrásnyelvre dolgoztuk, illetve dolgozzuk ki.

Az alapszótártól eltérő szókincs keresése. A forrásszövegből kiemeljük azokat a szavakat, amelyek a forrásnyelv szűk alapszótárában nincsenek benne. Ezzel kiszűrhetők a szakmai rövidítések, betűszavak és egyes egyszavas terminusok. Erre egy statisztikai vizsgálat is épül: a találatok közvetlen kollokációit is vizsgáljuk.

A terminusok belső morfoszintaktikai szerkezetének vizsgálata. Ez a módszer a többszavas terminusok megkeresésére alkalmas. Azt vizsgáljuk, hogy a többszavas terminusok belső morfoszintaktikai összetétele mutat-e olyan sajátosságokat, amelyek legalább részben megkülönböztetik őket a szöveg egyéb részeitől. Kísérletünkben legmélyebben ezt a módszert elemezzük.

A vizsgálat lényege: felírunk olyan morfoszintaktikai mintákat, amelyeknek megfelelő kifejezéseket ki szeretnénk emelni a forrásszövegből. Minden minta morfoszintaktikai kódok sorozata.

A keresés során a forrásszöveg szavait lemmatizáljuk, illetve elvégezzük morfológiai elemzésüket, így a szöveget lemmák, illetve morfoszintaktikai kódok sorozatává alakítjuk.

Az így átalakított szövegben egyszerű mintaillesztéssel keressük meg a prekoncepció alapján felírt mintákat, és kilistázzuk a mintáknak megfelelő szövegrészek felszíni formáját és lemmatizált alakját. A keresés során átugorjuk az úgynevezett tartalom nélküli szavakat (ezek listája szabadon szerkeszthető, így üres listával az ilyen szavak átugrása ki is kapcsolható). Ez a megoldás kezdetleges, sekély, lokális mondatelemzőnek is tekinthető (Jacquemin 2001).

A keresni kívánt minták listáját spekulatív módon is összeállíthatjuk, azonban kísérletünkben korpusznyelvészeti módszert is alkalmaztunk: ezt a módszer értékelésénél írom le.

A terminusok környezetének vizsgálata. Ez a módszer nominális terminusok megkeresésére alkalmas. A szövegben olyan egyszerű főnévi csoportokat keresünk, amelyek meghatározásszerű környezetben jelennek meg. Példák:

a terminus technicus egyértelműen szövegnyelvészeti jelenség

A fordító maga többnyire nem kompetens szakmai kommunikátor, hanem mediátor

Ezekben a mondatrészletekben a fókuszban levő főnévi csoportot kell észrevenni. Lényeges, hogy ezt nem a főnévi csoport belső szerkezetéből, hanem a környezetéből tudjuk meg.

A probléma általánosítása a szövegben előforduló definíciók általános, de sekély elemzése – arra jutottunk, hogy a környezetvizsgáló eljárásokat csak akként lehet megvalósítani.

Ezt a módszert jelenleg még nem alkalmazzuk, a környezetet leíró formalizmus kialakítása folyamatban van. A definíciók elemzésével kapcsolatos koncepciót azonban később leírtam a jelen módszer lehetséges továbbfejlesztéséről szóló részben.

Statisztikai módszerek.

Szokatlan gyakoriságú elemek keresése. Két módszert foglaltunk egybe: a köznyelvitől eltérő gyakoriságú szavak, illetve a köznyelvitől eltérő gyakoriságú szókettesek (kétszavas kollokációk) keresését.

A kísérleteinkből – és az irodalomból is – nyilvánvalónak tűnik, hogy pusztán a terminusjelöltek forrásszövegbeli gyakorisága nem igazolja (vagy cáfolja) egy kifejezés terminus voltát.

Mindkét esetben mérnünk kell egy referenciaadatot, amely nem más, mint az adott forrásnyelv egyes tartalmas szavainak előfordulási gyakorisága az adott

nyelv általános korpuszában. Az angol referenciaadatok kinyeréséhez a British National Corpus, a magyar adatokéhoz pedig a Nemzeti Szövegtár anyagát használjuk.

Ha nem alkalmazunk más módszert, ekkor meg kell mérnünk a forrásszöveg tartalmas szavainak, illetve a tartalmas szavakból alkotott szóketteseknek *(bigramoknak)* a relatív (a szöveg méretéhez viszonyított) gyakoriságát. Ezt a gyakoriságot kell összevetni a köznyelvi korpuszon mért (ugyancsak relatív) gyakorisággal. Kérdés még (nem vizsgáltuk), hogy milyen különbség számít szignifikánsnak, illetve a különbség küszöbértékét hogyan kell változtatni (ha egyáltalán kell) azon szavak és szókettesek esetén, amelyek a köznyelvi referenciakorpuszban ritkán fordulnak elő.

A terminusjelöltek listájára azokat a szavakat és szóketteseket vesszük fel, amelyek relatív gyakorisága a küszöbértéket meghaladó mértékben különbözik a köznyelvi korpuszban mért gyakoriságtól.

Amennyiben magyar – vagy bármely más, gazdag morfológiájú nyelven írt – szöveget vizsgálunk, nem a szóalakok, hanem a lemmák gyakoriságát kell kiszámítani, ahhoz pedig lemmatizáló programmal is rendelkezni kell az adott nyelvre.

Ennek a módszernek nagy hátránya, hogy a működtetéséhez nagy mennyiségű alapadat kell. Egy ilyen rendszer nemigen válhat végfelhasználói alkalmazás részévé: „legfeljebb" kiszolgálóoldali (internetes) szolgáltatásként működhet.

Asszociációs mértékek alkalmazása. Amikor többszavas terminusokat próbálunk megkeresni a szövegben, nem emelhetünk ki minden tartalmasszó-párt. Csak azok a szókettesek érdekesek, amelyek elemei (szavai) együtt nagyobb valószínűséggel fordulnak elő, mint külön-külön. (Villada 2005; Kis et al. 2004a; Kis et al. 2004b.)

A szókettesek ilyen tulajdonságait a korpusznyelvészet az ún. asszociációs mértékek szerint számítja ki. Ezek alapja a szókettesek elemeinek külön-külön, illetve együttesen mért gyakorisága. Erre különböző képleteket építenek, amelyek meghatározott szempontok szerint kiszámítják annak az eseménynek a valószínűségét, hogy a szókettes részei együtt inkább fordulnak elő a szövegben, mint külön-külön. (Kilgarriff, Tugwell 2001; Pedersen-Banerjee 2003)

A számítás eredménye olyan rangsorolt szóketteslista, amelynek az elején a legmagasabb affinitási pontszámmal rendelkező szókettes áll. Terminusjelöltként a lista első része jöhet szóba; meg kell határozni egy olyan minimális affinitási pontszámot, amelyet a szókettes el kell érnie ahhoz, hogy figyelembe vegyék.

Ez a módszer jól használható nagy korpuszok esetén, kisebb szövegekre alkalmazva azonban megbízhatatlanná válik. Így terminusjelöltek kivonatolására nemigen alkalmazható, hiszen a forrásszövegek nagysága rendszerint nem éri el a statisztikai kritikus tömeget. Határesetet jelentenek – és még jövőbeli vizs-

gálat tárgyát képezik – a teljes könyvet kitevő, néhány tízezer szavas forrásszövegek.

A módszerek összekapcsolása. A gépi szövegkivonatolási módszerek hatékonyságát (minőségét) két mérőszámmal mérjük:

- Fedés *(recall):* a szövegből az algoritmus által kiemelt, illetve a szövegben ténylegesen előforduló releváns nyelvi jelenségek aránya;
- Pontosság *(precision):* a szövegből az algoritmus által kiemelt ténylegesen releváns nyelvi jelenségek, illetve az összes kiemelt nyelvi jelenség aránya.

Különböző módszerek esetén e két szám is nyilván különböző lesz. Általában igaz, hogy az egyik javítása a másik romlását okozza, ezért vannak olyan eljárások, amelyek nagy fedést, de alacsony pontosságot nyújtanak, és vannak olyanok is, amelyek esetében a pontosság magas, de a fedés alacsony.

Ha több módszert alkalmazunk együtt, azokat érdemes előbb alkalmazni, amelyek nagy fedéssel működnek, és a listán olyan módszerek által érdemes szűkíteni, amelyek nagy pontosságot nyújtanak.

A fent felsorolt módszerek közül kimondottan nagy fedést és alacsony pontosságot nyújtanak a következők:

- A terminusok belső morfoszintaktikai szerkezetének vizsgálata
- Asszociációs mértékek alkalmazása

Potenciálisan nagy pontosságú módszerek a következők:

- A terminusok környezetének vizsgálata
- Szokatlan gyakoriságú elemek keresése

Ez azt jelenti, hogy az első két módszer által adott listát érdemes a második két módszer valamelyikének alkalmazásával szűrni: például megvizsgálhatjuk, hogy a mintaillesztéssel kapott elemek közül melyek fordulnak együtt elő valóban sokszor, illetve melyek azok, amelyek ténylegesen megjelennek terminusokra jellemző környezetben. Mindezek a kombinációk a kézi utómunka csökkentését szolgálják.

Induktív terminológiakeresés. Az előbbiekben abból indultunk ki, hogy nem áll a rendelkezésünkre semmilyen szótár a forrásszöveg tárgykörében. Azonban ez sokszor nincs így: bár ebben a tanulmányban nem foglalkozunk ezekkel a módszerekkel, a kutatási projektünknek része, hogy a számítógéppel a forrásszöveg és kiinduló szószedet alapján állítsuk elő a szövegben előforduló terminusok jegyzékét (Jacquemin 2001).

Ennek egyfelől alapja egy olyan program, amely egy szószedet forrásnyelvi tartalmának minden előfordulását megkeresi a szövegben. Ez különösen gazdag morfológiájú nyelvekben nehéz, ott lemmatizáló program körültekintő alkalmazására van szükség.

Ha ez rendelkezésünkre áll, kétféleképpen folytathatjuk a keresést:

- keressük a kiinduló szószedet forrásnyelvi kifejezéseinek kollokációit a korábban említett módszerek valamelyikével;
- keressük azokat a kifejezéseket, amelyek a kiinduló szószedet valamely kifejezésétől csak egy szóban térnek el.

A kiinduló szószedettől mélyebben is elvonatkoztathatunk: erről a kísérlet leírásában lesz szó.

Az első kísérlet[38]

Részletes kísérleteket a fentebb leírt eljárások közül kettővel (az alapszótártól eltérő szókincs keresése, a terminusok belső morfoszintaktikai szerkezetének vizsgálata) végeztem. Azt is vizsgáltam, hogy a többszavas terminusok milyen morfoszintaktikai kötöttségeket mutatnak az általában vett kollokációkhoz képest.

Az absztrakt minták előállítása. Evégett készítettünk egy olyan programot, amely egy szószedet többszavas kifejezéseinek minden előfordulását kigyűjti egy szakmai korpuszból, meghatározza azok morfoszintaktikai szerkezetét, és statisztikát készít mind az egyes terminusokról, mind pedig az absztrakt morfoszintaktikai mintákról – ahol a terminusokban levő szavak felszíni alakjait és lemmáit is eltávolítottuk, és csak a morfoszintaktikai kódok által alkotott mintákat tekintettük.

Kiinduló szószedetként a MorphoLogic számára elérhető informatikai szótárak angol és magyar többszavas címszavainak összességét használtuk. A minták kivonatolásához pedig a SZAK Kiadó által kialakított, informatikai szakszövegekből álló kétnyelvű párhuzamos korpuszt használtuk (vö. Kis Á.-Kis B. 2003).

- A szószedet 30 765 angol, 23 186 magyar többszavas kifejezést tartalmazott.
- A korpusz angol oldalának tömege kb. 1,2 millió szövegszó, a magyar oldalé kb. 1,6 millió szövegszó (az eltérésnek az az oka, hogy a korpusz eredeti magyar nyelvű műveket is tartalmaz).

A mintakivonatolás eredményéből egyszerűsített mintahalmazokat nyertünk. Ehhez 2 225 angol és 2 520 magyar mintát kellett megszűrnünk. Kétféle módszert alkalmaztunk:

- csak a 100-szor vagy gyakrabban előforduló mintákat használtuk fel;
- 1000 véletlenszerűen kiválasztott mintát manuálisan megszűrtünk, és az így kapott listát használtuk fel.

Az angol adatokra csak a második eljárást végeztük el.

Példák az ismert terminusokat kereső program által kiemelt mintákra

… angolul:

4037	[N]+[N]	dialog box	dialog+box	dialog[N]+box[N]
947	[ADJ]+[N]	hard disk	Hard+Disk	hard[ADJ]+disk[N]
880	[N]+[N]	check box	check+box	check[N]+box[N]
497	[N]+[N]	control panel	Control+Panel	control[N]+panel[N]
64	[ADJ]+[N]+[N]	direct cable connection	direct+cable+connection	…
18	[N]+[V][GER]	disk striping	Disk+striping	disk[N]+stripe[V][GER]

… magyarul:

985	[ADJ][NOM]+[N][NOM]	operációs rendszer	operációs+rendszer
320	[ADJ][NOM]+[N][ACC]	operációs rendszer	operációs+rendszert
69	[ADJ][NOM]+[N][INS]	operációs rendszer	operációs+rendszerrel
55	[ADJ][NOM]+[N][INE]	operációs rendszer	operációs+rendszerben
36	[ADJ][NOM]+[N][DAT]	operációs rendszer	operációs+rendszernek
26	[ADJ][NOM]+[N][ALL]	operációs rendszer	operációs+rendszerhez
24	[ADJ][NOM]+[N][ABL]	operációs rendszer	operációs+rendszertől

Új terminusjelöltek kiemelése az absztrakt minták alapján. A következő lépés az volt, hogy a korpusz elemzésével nyert mintákat kipróbáljuk: új terminusokat emeljünk ki új szövegekből.

Készítettünk egy olyan programot, amely az absztrakt minták alapján terminusjelölteket emel ki szövegekből, vagyis teljes mértékben megvalósítja a második szabályalapú algoritmust. Ez a program jelenleg angol és magyar nyelvű szövegek feldolgozására alkalmas.

Ennek a programnak paraméterül adtuk a fentebb említett absztrakt mintákat, s feldolgoztuk két-két informatikai szakkönyv teljes szövegét. A SZAK korpuszban a kísérlet idején egyik könyv sem szerepelt.

Számokban ez a következőt jelentette:

- Az angol szöveg terjedelme 338 215 szövegszó volt, ebből a program 25 705 mintát (13 869 különböző mintát) emelt ki;
- A magyar szöveg terjedelme 230 389 szövegszó volt, ebből a program 15 141 mintát (14 398 különböző mintát) emelt ki.

Példák a kiemelt mintákra

… angolul:

6	terminal service	Terminal services	[N]+[N][PL]
6	warning element	Warning element	[N]+[N]
6	worker process isolation mode	worker process isolation mode	[N]+[N]+[N]+[N]
6	XML parser	XML parser	[UNKNOWN]+[N]
4	server role	server roles	[N]+[N][PL]
4	shadow copy client	Shadow Copy Client	[ADJ]+[N]+[N]

… magyarul:

3	automatikus rendszer-helyreállítás	automatikus rendszer-helyreállítás	[ADJ][NOM]+[N]
3	hozzáférési jog	hozzáférési jogok	[ADJ][NOM]+[N][PL][NOM]
3	tartomány-nyilvántartó központ	tartomány-nyilvántartó központ	[N] +[N]
2	aktív tartalom	aktív tartalom	[N][NOM]+[N][NOM]
2	biztonsági házirend	biztonsági házirend	[N][NOM]+[N][NOM]
2	elérési út	elérési út	[N][NOM]+[N][NOM]

Módosított kísérlet: lépés a gyakorlati alkalmazások felé. A módszert kis módosítással felhasználtuk tényleges fordítás-előkészítési munkára is. Ezúttal spekulatív úton állítottunk elő mintasorozatokat, 1, 2, 3 és 4 szavas terminusok modellezésére.

Alkalmaztuk az első szabályalapú módszert is, vagyis kiemeltük a szövegekből azokat a szavakat, amelyek nem szerepeltek egy kb. 20 000 szavas angol alapszótárban. A két módszer által adott terminusjelölt-listát egyesítettük, s együttesen értékeltük ki.

Két angol nyelvű könyvet dolgoztunk fel, az első könyv terjedelme 100 963 szövegszó, a másodiké 74 626 szövegszó volt.

Az első kísérlet eredmények értékelése

Generált minták. A szótár és gyakorlókorpusz felhasználásával generált mintasorozattal kiemelt terminusjelölt-listákat a következőképpen értékeltük:

- Mindkét listából véletlenszerűen kiválasztottunk annyi elemet, amennyit elfogadható idő alatt manuálisan át lehetett vizsgálni. Ez 1 936 angol mintát (3 743 előfordulást), illetve 2 083 magyar mintát (2 107 előfordulást) jelentett.
- Két kutató egymástól függetlenül átvizsgálta a listát, és szavazott az egyes terminusjelöltekről. Csak azokat a terminusjelölteket tekintettük terminusnak, amelyeket mindkét kutató annak talált.

Ennek alapján az eredmények:

- 2 412 helyes angol minta (64,44%-os pontosság)
- 968 helyes magyar minta (45,94%-os pontosság)

Úgy találtuk, hogy a viszonylag gyenge eredmény oka a nagyon produktív szavak megjelenése a minták elején (ritkábban végén):

- angol: *new*, *all*, *other*, *same*, *such* stb.
- magyar: *elérésű*, *adott*, *alábbi* stb.

Ha automatikusan kiszűrjük a jelöltek közül azokat, amelyek a fenti szavak valamelyikével kezdődnek, az eredmény ugrásszerűen javul:

- angol: 77,08%-os pontosság
- magyar: 67,08%-os pontosság

Ez azonban még tovább is javítható lenne, például, ha találnánk megoldást a szószedetes kereséssel kapott minták automatikus utószűrésére, vagy bevezetnénk a produktív szavak automatikus felismerését.

A kivonatolási folyamat feltételezésünk szerint tovább javul, ha a morfoszintaktikai kivonatoló eljárás kimenetét tovább szűrjük statisztikai módszerrel, asszociációs mértékek felhasználásával. Bár a pillanatnyilag rendelkezésre álló rendszerben mindkét módszer meg van valósítva, részletes kísérleteket még nem végeztünk.

Kivonatolás spekulatív mintákkal. A spekulatív minták alkalmazása sok produktív szót kizárt, elviekben tehát jobb eredményt kellett kapnunk. A feladat azonban ezúttal nem a terminusjelöltek egyszerű pontozása volt, hanem két könyv fordításának tényleges terminológiai előkészítése.

A terminusjelöltek listájának szűrését ezúttal is manuálisan végeztük, de ezúttal minden (különböző) mintát megvizsgáltunk (a minták statisztikai listáját vizsgáltuk át). Az eredmények az 5.2. táblázatban láthatók.

	Terjedelem (szövegszó)	*Előfordulások*	*Különböző minták*	*Elfogadott minták*	*Pontosság*
1. könyv	100 963	25860	5523	1595	28,88%
2. könyv	74 626	14330	6535	2275	34,81%

5.2. táblázat. A spekulatív mintákkal kivonatolt terminusjelölt-listák számszerű jellemzői

Ez az eredmény szemlátomást sokkal rosszabb, mint az automatikusan generált minták esetén. Azonban a feltételek is mások:

- Mindkét lista sok tulajdonnevet tartalmazott, amelyeket bizonyos szempontok alapján terminusként is lehetne kezelni, itt azonban csak zavaró lett volna, ezért töröltük a listáról.
- A szószedetben a szócikkek számát igyekeztünk 2000 körül vagy az alatt tartani, mert a fordításokat egyelőre még manuálisan adjuk hozzá a szószedethez, és ez az a méret, amely – figyelembe véve a könyvek fordításának határidejét – még elfogadható idő alatt feldolgozható.

A pontosságértékből nem látszik az a hatékonyságnövekedés, amelyet e módszer jelent. Ha ezt az eljárást a terminológus számítógép nélkül hajtja végre, akkor végig kell olvasnia az 500, illetve majdnem 400 oldalas könyveket, és kézzel meg kell jelölnie minden terminust – gyors munka esetén 25 oldalt lehet átvizsgálni egy óra alatt, így egy 500 oldalas kötet kivonatolása 20 órát igényel. Ehelyett mindkét esetben körülbelül 2 perces programfutás és egy kb. 3 órás utószűrési fázis következett, amely lényegesen kevesebb, mint a könyvek ember általi végigolvasása.

A második kísérlet

A terminuskivonatoló program bővítése. A fejlesztés egy későbbi fázisában a programot kiegészítettük induktív funkciókkal is. A konkrét könyvkiadói alkalmazásban ugyanis szükséges volt kihasználni a korábban létrehozott glosszáriumok tartalmát is.

Az új program induktivitása valójában azt jelenti, hogy a statisztikai és a szabályalapú kivonatoló algoritmusok által visszaadott terminusjelölt-listát pontosabban értékeljük a kiindulási szószedet segítségével. Az algoritmusok ugyanis 100%-os fedésre vannak beállítva.

A kivonatolás után a kapott jelöltek automatikus értékelése következik. A szabályalapú algoritmusok minden jelölthöz hozzárendelnek egy kiinduló pontszámot, az utóértékelő funkció pedig ezt módosítja:

- jelentősen megnöveli, ha a lemmatizált terminusjelölt pontosan egyezik az indukciós szótár valamelyik elemével;
- kisebb mértékben növeli, ha a lemmatizált terminusjelölt egy szó híján egyezik az indukciós szótár valamelyik elemével;
- a 3 szavas vagy hosszabb terminusjelöltek esetén kis mértékben növeli, ha a lemmatizált terminusjelölt két szó híján egyezik az indukciós szótár valamelyik elemével.

Az utóértékelő modul a terminusjelöltek gyakorisága alapján is módosítja a pontszámot. A gyakoriság szerint a terminusjelölteket három kategóriába lehet sorolni, és mindegyik kategóriához hozzá lehet rendelni egy pontszámnövekményt.

A pontozási szabályok részletesen szabályozhatók, így a programmal az általunk végzettnél kiterjedtebb kísérleteket is el lehet végezni.

Az új program az indukciós szótárt nemcsak pontozáshoz használja; a kimenetbe beírja a szótárban talált CNy megfelelőt is.

A második kísérlet eredményei. Az új program spekulatív alapbeállításaival elvégeztük a terminuskivonatolást két könyv szövegén. Az eredményeket az 5.3. táblázat mutatja:

1. könyv:

Pontszám	*Jelöltek száma*	*Jelöltek aránya*	*Pontosság*
120 fölött	315	3,51%	89,52%
100 fölött	1286	14,33%	32,27%
90 fölött	2902	32,33%	31,63%
Összesen	8976	100,00%	19,47%

(folytatás a következő oldalon)

2. könyv:

Pontszám	*Jelöltek száma*	*Jelöltek aránya*	*Pontosság*
120 fölött	462	3,54%	74,68%
100 fölött	1217	9,32%	47,90%
90 fölött	2561	19,61%	37,25%
Összesen	13060	100,00%	21,87%

5.3. táblázat: A második kísérlet eredményei

Látható, hogy amennyiben az összes terminusjelölt körében vizsgáljuk a pontosságot, nem kaptunk jobb eredményeket, mint az első kísérletben. A pontszámok azonban hatékonyan kategóriákra osztják a jelöltlistát, és látszik, hogy a pontszám valóban megbízhatósági mérték: meg tudunk határozni olyan pontszámot, amely mellett jelöltek túlnyomó többsége helyes.

További fejlesztések

A szövegbeli definíciók elemzése. Korábban említettem, hogy a környezetvizsgáló eljárások a szövegben előforduló, definíciószerű mondatok elemzésére épülnek majd. Ez arra a hipotézisre épül, hogy a terminológia előfordulása meghatározott kontextusban valószínűbb: a terminológiai helyzet ugyanis elsősorban nem a terminus struktúrájából, hanem a terminus környezetének struktúrájából következik. A lehetséges kontextusok a címek, a definíciók és a hozzájuk hasonló fogalombevezető mondatok (a magyarban ilyenek például a 'nevezzük' igére végződő, a terminust részeshatátozó-esetben tartalmazó mondatok).

A környezetvizsgáló eljárásoknak két feladatuk van:

- a definíciók és fogalombevezető mondatok felismerése és
- az ilyen mondatok felbontása elemeikre.

Az eddigiekben elvégeztem néhány definíció, illetve definíciószerű szövegbeli mondat kézi elemzését, szigorúan a gépi főnévicsoport-kereső algoritmust szimulálva. A definíciók elemeit az arisztotelészi genus proximum–differentia specifica fogalomrendszer szerint azonosítom. A definíciók elemeit az alábbi példákban különböző aláhúzások jelzik:

- definiendum vagy terminus
- genus proximum
- differentia specifica

Szótári definíciók:

opportunity cost: the loss of other alternatives when one alternative is chosen.[39]

egér: Az a síkban mozgatható eszköz, amellyel a számítógépen a kurzor helyzete változtatható.[40]

többes indítású rendszer: olyan számítógép-rendszer, amelyen egyszerre több operációs rendszer van telepítve, s a számítógép indításakor indításfelügyelő program segítségével ki kell választani, melyiket kívánjuk betölteni.[41]

Szövegbeli definíciók:

A teljes morfológiai rendszer gyakorlatilag *nem más, mint* több különböző morfématár együttese.[42]

A grouping of 32 bits *is called* a word, as shown here:[43]

Az utasítás *lehet* egyetlen parancs vagy több parancs lánca vagy zárójeles csoportja.[44]

Feature-set control in the early part of a project *consists* primarily *of* not putting unnecessary features into the product in the first place.[45]

Látható (és csöppet sem meglepő), hogy a szövegbeli, definíciószerű mondatok szerkezete sokkal kevésbé kötött, mint a szótári definícióké. Ezért Mihaltz (2004) megközelítése a genus proximum szótári definícióból való kiemelésére valószínűleg csak korlátozottan lesz használható.

A fenti példák dőlt betűkkel kiemelve jelölik azokat a szavakat, amelyekből felismerhető a definíciószerűség, és amelyek alapján – nyelvfüggően – meghatározható, hogy a terminusként azonosítható főnévi csoportok hol találhatók. A példákban szaggatottan aláhúzott – tulajdonképpen genusként megjelölt – főnévi csoportokról sokkal kisebb biztonsággal jelenthetjük ki, hogy terminusok.

Végfelhasználói alkalmazás fejlesztése. Az új program a környezetvizsgáló eljárások nélkül is alkalmas arra, hogy végfelhasználói alkalmazás részévé váljon, amely fordítástámogató eszközökbe ágyazva vagy azok mellett működve fordítókat, fordítócsoportokat segít a fordítások terminológiai előkészítésében.

A szövegszinkronizáló programokhoz hasonlóan itt is sokat számít a kezelőfelület (ami ehhez a programhoz még nem készült): az utószűrésre fordított idő jelentősen csökkenthető, ha a felhasználói felületen például egy művelettel lehet törlési szabályokat beállítani vagy több jelöltet törlésre kijelölni.

Kétnyelvű terminuskivonatolás párhuzamos korpuszból (fordítómemóriából). A terminológiakivonatolásnak az előbbiekben ismertetett eljárásai csak a forrásnyelvi szövegben megjelenő terminusok összegyűjtésére alkalmasak.

A terminológiai előkészítésnek azonban része a terminusok fordításának meghatározása is. Ennek triviális módja a meglevő szótárak felhasználása, s valóban szükséges, hogy a fordítások automatikus vagy legalább félautomatikus meghatározására minden rendelkezésre álló erőforrást felhasználjunk.

A párhuzamos korpuszok, illetve azok elégtelen volta esetén a forrásnyelvi korpuszok, olyan esetekben segíthetnek, amikor adott terminusok nem állnak rendelkezésre szótárban vagy terminológiai adatbázisban.

A nemzetközi szakirodalomban számos eljárás olvasható két- vagy többnyelvű terminuskivonatolásról. Blank (2000) a párhuzamos korpusz egynyelvű

részkorpuszain egymástól függetlenül végez terminuskivonatolást, majd a kapott jelölteket statisztikai eljárásokkal egymáshoz igazítja. Choueka et al. (2000) a párhuzamos korpuszok szószintű szinkronizálását javasolja. Callison-Burch et al. (2005) ezzel szemben a FNy terminusok előfordulásaihoz keres statisztikai módszerekkel CNy megfelelőket. Pohl (2006) főnévicsoport-kiemelést *(NP chunking)* alkalmazó módszere pedig a mondat- vagy bekezdésszinten szinkronizált párhuzamos korpuszban előforduló főnévi csoportokat igazítja egymáshoz.

A fenti módszerek azonban két probléma miatt nem alkalmazhatók közvetlenül a fordítástechnológiában:

(1) Rendkívül nagy méretű korpuszt igényelnek: Blank (2000) például nyelvenként 12 millió szövegszavas korpuszon kapott értékelhető eredményeket. Mivel a terminológiát csak azonos tárgykörbe tartozó, adott esetben azonos forrásból származó szövegeken lehet értelmezni, nemigen találunk olyan fordítót vagy fordítással foglalkozó szervezetet, amelynek birtokában lehetne több millió szövegszavas, tárgykör-specifikus párhuzamos korpusz.

(2) Nem törekszenek a FNy szöveg teljes terminológiai vázának leképezésére. Ez pedig, mint korábban láttuk, követelmény a fordítási projektekhez kapcsolódó terminológiaalkotással szemben. Ennek az az oka, hogy miközben az alkalmazott szabályalapú vagy statisztikai módszerek nem nyújtanak teljes fedést és nagy pontosságot sem, az irodalomban leírt módszerek nem számolnak esetleges emberi utófeldolgozással – vagyis az emberi munka megkönnyítése helyett annak teljes automatizálására helyezik a hangsúlyt.

A fentiek miatt az irodalmi módszerekre építve olyan alternatív utat keresünk, amely figyelembe veszi, hogy a kivonatolandó FNy szöveg néhány ezer szövegszóból, a segítségül hívható párhuzamos korpusz pedig legfeljebb néhány százezer szövegszóból áll. Emellett pedig a kivonatolási folyamat megtervezésekor számolunk az emberi utófeldolgozási munkával is (lásd az előző oldalt a végfelhasználói alkalmazásról!).

Alább vázolok néhány módszert, amellyel az értekezés írása idején kísérletezünk.

Ha adott terminus célnyelvi megfelelőjét keressük, általában statisztikai eljárásokat alkalmazunk. A következőkben kétféleképpen járhatunk el, attól függően, hogy a rendelkezésünkre álló korpusz szinkronizálva van-e vagy nincs.

Ha a párhuzamos korpusz legalább bekezdésszinten szinkronizálva van (de jobb a mondatszinkronizálás), akkor elegendő azokat a szegmentumokat megkeresnünk, amelyekben a forrásnyelvi terminus előfordul. Erre a fordítómemóriák konkordanciafunkciója jó lehetőséget ad, de amennyiben magunk fejlesztünk ilyen szolgáltatást, programot sem nehéz e célra írni – ha a szinkronizáló modul már rendelkezésre áll.

A megtalált szegmentumok célnyelvi megfelelőiben azt a kollokációt vagy szót keressük, amely lehetőleg az összes vagy majdnem az összes, a FNy termi-

nust tartalmazó szegmentumban előfordul. Ha a terminus kevés helyen fordul elő, vagy az őt tartalmazó forrásnyelvi mondatok nagyon hasonlítanak (esetleg egyformák), akkor nagyon sok jelölt lesz. Ez abban az esetben probléma, ha automatikus eljárást konstruálunk a megfelelők megkeresésére: ha a terminológus maga nézi át a megtalált forrásszegmentumokhoz tartozó célszegmentumokat – mert pl. fordítómemória konkordanciaszolgáltatását használja –, akkor valószínűleg rövid idő alatt ki tudja választani terminus célnyelvi megfelelőjét.

Ha a párhuzamos korpusz nincs szinkronizálva, akkor a forrásnyelvi terminus gyakoriságára és előfordulásainak szövegbeli eloszlására építhetünk. Ekkor is jól járunk, ha legalább a szövegek szegmentálását elvégezzük, de a konkrét szegmentumok megfeleltetése helyett (az volna a szinkronizálás) a szegmentumoknak a szöveg egészéhez viszonyított (százalékos) pozícióját tekintjük – így összehasonlítható eredményeket kapunk.

Ha automatikus eljárást konstruálunk, mindig fennáll annak a veszélye, hogy túl sok célnyelvi jelöltet kapunk. Ha egy terminus csak egyszer fordul elő a forrásszövegben, a jelöltek között az első körben megkapjuk a célszövegből a teljes *hapax legomenát*, a második körben (a pozíciók vizsgálata közben) pedig a valódi célnyelvi fordítás környezetében levő valamennyi szót és kollokációt. Ilyen esetben az automatikus művelet egyrészt nagyon sokáig tart, másrészt pedig használhatatlan lesz az eredmény. Sokan döntenek ezért úgy, hogy az ilyen vizsgálatoknál kihagyják a *hapax legomenát* (az egyszer előforduló jelölteket).

5.3. A fordítástechnológia és a lexikográfia

A fordítástechnológiában a terminológiakezelés hagyományosan azt jelenti, hogy a fordításhoz esetleg szükséges terminusokat terminológiai adatbázisban soroljuk fel, s az adatbázist elektronikus szótárként használjuk a fordítás közben (Austermühl 2001:107). Ezt az 5.1. fejezetben ismertetett terminológiai technológia túlhaladta már, azonban a legtöbb fordító ma is a nyomtatott vagy az elektronikus szótárakra alapozza a terminológiai kutatást.

A készen elérhető terminológiai adatbázisok nincsenek feltétlenül közvetlen kapcsolatban a konkrét fordítási feladattal: a legtöbbször adott tárgykör, illetve azon belül adott műfaj, esetleg dokumentumcsoport körére kidolgozott normatív terminológiával találkozunk, amely kiindulásként szolgál a későbbi, hasonló tárgykörhöz, műfajhoz, dokumentumcsoporthoz tartozó forrásszövegek fordítása esetén.

A fordítás és a szótárak kölcsönhatása

A fordító az esetek döntő többségében szótárakon keresztül kerül kapcsolatba a terminológiával. Bár a konzisztens szakmai kommunikációra való törekvés megnyilvánul abban, hogy a terminológiát – nyelvpolitikai és gazdasági okokból – igyekeznek szabványosítani (pl. Rey 1993:176-180 vagy Sager 1990:118-

120), a terminológiának sem a szabvány, sem a szótár nem elsődleges forrása. Ennek egész egyszerűen az az oka, hogy a terminológia szorosan kapcsolódik a mögötte levő, valós műszaki vagy más szakmai fogalmakhoz és objektumokhoz – amelyek pedig elsősorban a kutatás–fejlesztés során keletkeznek. Ezért, amint az 5.1. fejezetben is említem, a terminológia elsődleges forrásai a kutatási beszámolók és más szakmai írások.

A fordítás, a terminológia és a szakmai nyelvhasználat ezért folyamatos kölcsönhatásban állnak egymással. Ennek legfőbb elemei:

(1) A kutatás-fejlesztés során új dolgok, fogalmak jönnek létre. Az adott kutatás-fejlesztés eredeti – vagy a terület „hivatalosnak" tekintett – nyelvén ennek során új terminológia keletkezik.

(2) A fordítások során adott területhez a fordítás célnyelvén másodlagos terminológia keletkezik.

(3) A szabványosítási és egységesítési folyamatok során az (1) vagy (2) során létrehozott terminológia szótárba vagy terminológiai adatbázisba kerül.

(4) A szabványos, szótárba vagy terminológiai adatbázisba foglalt terminológiát az adott területen, az adott célnyelvre irányuló fordítások során már nem kell újra kidolgozni, a szótár ekkor már forrásként funkcionál.

Mondhatjuk tehát, hogy a fordítás egyszerre forrása és felhasználója a terminológiának. Álljon itt erre két hazai példa:

- Az Európai Unió jogharmonizációs fordításai során keletkezett terminológia 2004 végén szótárba került (Várnai-Számadó 2004), és a továbbiakban normatív terminológiaforrás az uniós szövegekkel foglalkozó fordítóknak. Mivel ennek forrása az Igazságügyminisztérium Fordításkoordináló Egysége, a szótár korpusztervezési befolyása – függetlenül a tartalmának minőségétől – rendkívül erős.
- A SZAK Kiadó a 2000-es évek elejétől korpuszba rendezi az általa megjelentetett informatikai műveket, s gyűjti a fordításokhoz kialakított – részben új, részben pedig hardver- és szoftvergyártóktól átvett – terminológiát. 2003-ban és 2005-ben ezt a terminológiát szótárban megjelentette (Kis 2005), amelyet így az informatikai szövegek fordítói segítségként használhatnak fel. Mivel azonban ennek forrása egy kis címszámú könyvkiadó, a szótár korpusztervezési hatása korlátozott, pontosabban attól függ, hogy az adott tárgykör fordítói milyen tömegben ismerik meg és fogadják el.

A gyorsan fejlődő szakterületeken a terminológia igen gyorsan elavul. A terminológiai szótár az adott szakterület nyelvhasználatának – szociolektusának – pillanatfelvételét, szinkrón modelljét rögzíti. Így az idő – nem évek, hanem hónapok – előrehaladtával a szótár tartalma egyre kevésbé lesz autentikus forrása a szakterület terminológiájának, s így a szakmai fordításokhoz is egyre kevésbé lesz használható. Austermühl (2001) munkája a fordítás számítógépes segédeszközeiről meg sem említi a szótárakat.

A fordítás és a számítógépes lexikográfia

Számos példát hozhatunk tehát arra, hogy a fordítási munka során keletkező terminológia szótárba kerül. Ezért azt mondhatjuk, hogy kétféle lexikográfia létezik:

(1) Proaktív: szótárírás konkrét fordítási feladat motivációja nélkül, nyelvoktatási, általános vagy szakmai kommunikációs céllal, korpuszalapon vagy spekulatív módon. Az általános szótárak szinte mindig így készülnek.

(2) Retrospektív: konkrét fordítási munka vagy munkák együttesének során kialakított terminológia szótárba rendezése és publikálása. Ezt általában fordítással foglalkozó szervezetek végzik.

A terminológia kapcsán szinte mindig a retrospektív lexikográfia kerül előtérbe, hiszen ez azt jelenti, hogy fordítási szószedetek – terminológiai adatbázisok – tartalmát jelentetik meg önálló nyomtatott és/vagy elektronikus szótár formájában.

Mivel e dolgozatnak nem tárgya lexikográfia, itt csak annyit említek, hogy a terminológiai adatbázisban meglevő szócikkek egyértelműen és könnyen alakíthatók XML-formátumra (már ha nem rögtön ebben a formátumban jelennek meg), amelyek pedig automatikusan formázhatók, tördelhetők. Így, ha konzisztens módon felépített terminológiai adatbázisból indul a szótárkészítés, egy nagyobb szótár átfutási ideje is rendkívül rövid lehet (akár 2 hónapnál is kevesebb, mint a korábban említett európai uniós szóanyag esetében is). (Lásd még: Kis B.-Kis Á. 2003)

Sok fordító ír terminológiát munkája közben. Ehhez általában közönséges szövegszerkesztő vagy táblázatkezelő programot használ. Azonban fontos lenne előmozdítani a terminológiakezelő programok használatát – és a jelenlegieknél egyszerűbben kezelhető terminológiakezelő programok fejlesztését –, ugyanis ezek segítségével külön odafigyelés nélkül lehet egységes szerkezetű szócikkeket létrehozni.

Summary in English

Introduction

Recently, translation has been subject to a significant paradigm shift. While earlier it was considered as individual creative work, nowadays it is rather difficult to find a translation task that can be accomplished by one individual translator. Today translation, like many other creative activities, is performed by teams, and because the deadlines have become tighter, many organizational and technical means needed to be deployed so that the tasks could be performed in time.

The economic significance of translation and the related technology is indicated by the fact that the professional community organises numerous conferences on the subject. An example is Localization World, or St Jerome's Day in Hungary, organized by the largest association of translation companies – and one must also mention the events organized by the largest international community, Proz.com. However, little academic literature is available on translation technology. There is significant research activity on related subjects such as machine translation, corpus linguistics, translation training or language policy: the bibliography at the end of this thesis includes both classic and contemporary works. However, no such writings are available on the technical and technological aspects of translation. The sources considered the most important in the field (Esselink 2001, Austermühl 2001) are technical descriptions or textbooks rather than academic monographs.

The Purpose and Structure of the Thesis

I regard this thesis as a summary of the field. Its main purpose is to define translation technology as a distinct field within applied linguistics. It is quite obvious that translation technology is closely related to translation studies, natural language processing and corpus linguistics. However, it is equally important to mention its relation to sociolinguistics and language policy – because it is precisely translation technology that makes it possible to accomplish today's translation tasks of increasing size and tighter deadlines.

The thesis aims at fulfilling this purpose by describing or demonstrating research in translation technology. The research activities are systematically aligned with various fields of applied linguistics. The first chapter defines translation technology as a field of research; in the second, its relation to language policy is described (cf. Szépe 2001; Szabari 1996; Horváth 2002). I am convinced that education cannot be separated from any fields of applied linguistics; the chapter on language policy also includes a description of teaching various aspects of translation technology (Kis, B. 2004, and Drugan 2004). The third chap-

ter undertakes a seemingly easy task, namely, the description of the connection between translation technology and translation studies. In the fourth chapter, corpus linguistics and natural language processing is applied to translation technology, while the fifth chapter deals with the problems of terminology in translation.

1 Defining Translation Technology

1.1 Translation Technology as a Field of Research

The purpose of translation as a business activity is producing a target-language text on the basis of the source text received. Individual translation is becoming rare: translation is increasingly performed as group work, aided by technical devices. This follows from the size of the average translation task and the time available to accomplish it. By general agreement, it can be determined when a target-language text can be considered the translation of a given source-language text. Translation teams produce the target-language text according to (often informal) rules determined by both this general agreement and the constraints of group work.

Translation is thus regarded as a technical – manufacturing – activity, where products are created by means of specific devices, and following specific procedures and rules. This is technology. Translation is a technical activity to the extent that even standards apply to it: UNI 10574 (Italian), Önorm D 1200 és D 1201 (Austrian), DIN 2345 (German), Taalmerk (Dutch), ISO 12616 (international), EN-15038 (European) (cf. Arevalillo 2007).

When the translation process is viewed from the aspect of translation technology, we can distinguish between micro-strategies and macro-strategies used by translators. The micro-strategy of translation applies to the atomic operation, namely, translating one segment. A segment is a limited linguistic structure, in most cases, one sentence. Micro-strategy defines the process of the translator moving from the source-language segment to producing the translation, taking into account the assistance from various resources.

The macro-strategy of translation defines methodologies describing processes consisting of the above-mentioned atomic operations. These processes include the preparation of the work, i.e. dividing the source text into documents, and the documents into segments. Then translated documents are formed from the translations of the individual segments; quality assurance is performed; finally, the entire target-language material is composed from the translations of the documents.

Translation technology is in interaction with various fields of applied linguistics, while it is also viewed as one of the potential research topics of translation studies.

As a field of research, translation technology is in systematic connection with multiple fields of applied linguistics:

(5) language policy (sociolinguistics): the existence and proper execution of translation relates to linguistic rights, and, in many cases, it is prescribed by law. Considering the current demand for translation, and the large size of such translation tasks, their proper accomplishment is not possible without the means and processes of translation technology. Thus it receives priority in language planning, including both status and corpus planning.

(6) translation studies, in several respects: translation technology influences the translation process, and, at the same time, facilitates research on certain aspects of translation. This is achieved by the creation of parallel corpora during translation. On the other hand, due to the well-defined processes, quality assurance can be observed, and new models can be set up for translation equivalence.

(7) corpus linguistics and computational linguistics: translation memories and term bases resulting from the translation process can serve as linguistic materials for research on language analysis and machine translation. In addition, the comparison of texts before and after proofreading facilitates the automation of correcting translations.

(8) terminology studies:[1] correct and consistent terminology use is a substantial feature of technical translation. Translation technology is almost exclusively applied to technical translation. Preparation, use and quality control of terminology has special importance; even more so because in a target language, many terms are introduced through translation. Therefore, the translation technology workflows usually include a terminology workflow as well.

Because translation technology has also been defined as a technical field, it also has interdisciplinary connections to areas of technology:

- computer science, including human language technologies: while parallel corpora and the integration of machine translation are popular research topics, translation technology systems require the storage and high-performance processing of large amounts of linguistic data, and the efficient exploitation of existing parallel corpora. Development of such systems presents a demand for non-trivial data models and searching algorithms.

[1] In principle, the „-logy" suffix in the term „terminology" refers to the scientific nature of the field. However, the common interpretation of terminology is rather the set of terms of a specific domain. The reason for adding the word term 'studies' is that I found it important to distinguish between the field of research and the linguistic means. Chapter 5 includes a detailed explanation of the related meta-terminology.

- process control and project planning: a translation technology system requires a well-defined workflow. Translation tasks nowadays require translation organizations to create a complex project.

1.2 Translation Technology and Machine Translation

Machine translation and machine-aided translation (or computer-assisted translation, abbreviated as CAT) has different purposes. The distribution of tasks between the two paradigms stems from the view that machine translation has no use in translation as such. This opinion was induced in researchers and the society by the ALPAC report (Pierce, Carroll et al. 1966:32). As early as in the early 1980s, Martin Kay of XEROX made essential contribution to placing man and machine to their proper places in translation (Kay 1980).

The distinction originates in a substantial difference: while machine translation is fast and automatic, and produces translations of poor linguistic quality that is often only partly legible, machine-aided human translation is human translation by its essence, so its quality can potentially reach the best possible human translation. Although machine-aided human translation requires significantly more time and effort than producing automatic machine translation, it is far quicker than human translation without assistance.

The quality gap between fully automatic machine translation and machine-aided human translation implies that the two procedures are applied at different stages of the communication process. While automatic machine translation facilitates the reception (comprehension) of the communication, machine-aided human translation is instrumental in the transmission (production) of communication.

It is also implied that all computational systems involved in language production or translation assume that human intelligence is incomparably superior; therefore human output is always preferred to machine output in terms of quality. The often implied design philosophy behind today's computer systems is based on the belief that human output can always be considered superior without prior examination. However, as Melby (1995) observes, „[...] bad human translation is interesting because it was most likely done by a human yet in a manner similar to the way computers translate".

2 The Role and Impact of Translation Technology in Language Policy

This chapter summarizes the impact of translation technology on language policy. It describes how certain aspects of language policy motivate the creation of the field. The changes in the translation community are also covered, as well as the effect of translation technology on language planning.

Education in computer-assisted translation and translation technology is also addressed. I am convinced that education is closely connected with language planning because in this respect it can be considered as the systematic transfer and enforcement of social and economic changes.

2.1 The Need for Translation Technology

Most social groups and organisations have a global demand for communication. Global communication requires local communication, that is, translation. Translation has limited resources. However, there is an increasing demand for immediate or real-time communication, and, accordingly, immediate translation. For the time being, as only humans can produce translations of proper quality, the efficiency of human translation must be improved. Part of this effort is using automatic machine translation: because of the demand, it inevitably has a *raison d'être,* although its present state raises doubts about its usefulness.

2.2 Social and Economic Aspects of Technologized Translation

The new environment of translation requires team translation. In addition to computer assistance, it gives rise to important organizational issues that deserve detailed observation.

If the source material is of a large amount and the deadline is tight, it is obvious to perform the task in a parallel manner. If one person is incapable of completing the translation in due course, a translation team must be employed. It is shown in the thesis that once the translation is performed in a parallel manner, computer assistance cannot be avoided. This is justified by the requirement of consistency.

As a result, translation work becomes increasingly technologized: in team translation, the individual, un-coordinated translation becomes a well-defined process consisting of three distinct phases (preparation–execution–post-processing), with well-defined distinct tasks in each phase.

Team translation demands new means and new skills. The centre of team translation is an organization with the purpose of carrying out workflow management, and the creation and distribution of translation resources. A translator working in a team is always working in a network, and is in continuous communication with teammates.

2.3 The Role of Translation Technology in Status Planning

In this section and the next one, I propose to study translation technology in the light of the taxonomy introduced by Einar Haugen (1983), namely, that of status planning and corpus planning.

The issue of status planning is a rather straightforward one: as long as translation is provided into a language or a dialect in a subject field or an international organization, respecting the nature of the target culture, the equal status of the language and the community in question can be maintained with respect to the particular subject field or within the international organization.

Local or decentralized status planning also includes terminology creation in translation. Although Ádám Kis views this rather as a corpus planning process (cf. Kis, Á.-Kis, B. 2004), corpus planning can always be considered as a means of the implementation of status planning. Recognizing this, experts in Hungary established the Terminology Council of the Hungarian Language (MATT), involving governmental and non-governmental organizations, enterprises and private individuals, also with the endorsement of the Hungarian UNESCO Committee.

2.4 The Role of Translation Technology in Corpus Planning

Applying translation technology is also a matter of corpus planning because the conceptual system described in the source language is often transferred to the target culture exclusively through translation. As a consequence, translation quality assurance and terminology planning becomes significant because the language use of the subject field in which the translation is done is primarily affected by the quality of terminology and translation. This circumstance imposes the following tasks on those dealing with translation and the particular subject field:

- Systematic and co-ordinated terminology creation independently from translation
- Systematic and co-ordinated terminology creation and use in translation
- Production of translations of high level of consistency and good linguistic quality (legibility).

In short, the double task of status planning and corpus planning is to ensure the appropriate supply of translation services on the one hand, and the terminological basis and the translation consistency on the other. If source-language documents are created at a high rate, and, at the same time, the available supply of translation services is limited; this task cannot be accomplished without the help of technology (cf. Rey 1995:167-180). Currently available technological means might prove insufficient for this purpose. However, further tools exist, many of them in the experimental phase, which could provide additional assistance.

2.5 Teaching Translation Technology

Language planning activities related to translation technology must include education of macro- and micro-strategies of translation technology – as translation technology cannot be employed without translators and other co-workers who are aware of the processes and have sufficient skills to use the necessary equipment. This includes the use of technical means of translation in the micro-strategy, the organization of translation tasks, and the technical services behind the translation workflow (i.e. the macro-strategy).

Translators and other co-workers need the following skills:

- production skills: to preserve the competitiveness of their translation performance and availability (Austermühl 2001);
- network communication skills: to eliminate relative isolation, and to be able to work in (virtual) teams;
- maintenance skills: to maintain the operation of their own IT infrastructure even when no external assistance is available.

Over the past few years, I have developed a course of translation technology for translation students. In the thesis, this course is described from three aspects: (1) the existing knowledge and skills of students; (2) the priorities of the syllabus; (3) the methodology.

The methodology prioritizes the individual work of students. In class, students almost always perform individual exercises or group work. Teacher contact is necessary to provide the students with easy access to the self-paced learning materials. Exercises and other materials are available over the Internet, and are prepared in such a way as to be suitable for self-study as well. The materials, although they were started as an individual project, are now being developed in a wiki structure by a group of translation technology teachers from various institutions.

3 The Interaction of Translation Technology and Translation Studies

Translation technology is useful to translation studies both as a resource and an extension of the research domain, as Chapter 3 shows.

To establish the connection between translation technology and translation studies, one must formulate those questions of translation studies that are instrumental in the interaction of the two fields. A preliminary formulation of these questions can be as follows:

- What are the cognitive processes of translation?
- What is translation equivalence, and how can it be studied?

- What are the grammatical, semantic and other linguistic characteristics of translated texts? How are these characteristics related to the properties of the source texts?

Further, Chapter 3 makes the following observations:

1. Translation technology is subject to research on translation studies. As translation studies are interested in both the cognitive processes of translation, and the linguistic properties of target-language texts, they consequently need to assess the effects on these imposed by the employment of translation technology. Because I was not involved in such research, I could only outline the possible research methodologies by which translation scholars can achieve results in the subject.

2. Resources created by the use of translation technology provide means for investigations into the translation process, the translation equivalence, and the effect of the source language on translated texts.

3. The use of translation technology resulted in a new equivalence model. In this model, the only condition for a translation to be equivalent is acceptance for publication by an editor that adheres to a certain technological procedure. This procedure can be summarized as the preservation of all stages of the target text, including the one immediately after translation and the one ready for publishing.

3.1 Equivalence and Quality

We have no technical means to assess the total (mental or semantic) equivalence in itself, researchers usually attempt to grasp translation equivalence at various levels (see, for example: Catford 1965, Komissarov 1990). Once computers are involved in the process, the particular way of thinking proposes the decomposition of the communication into smaller units where equivalence can be formally described and evaluated – if the equivalence of the entire communication is unsuitable for proper investigation. The formal description of equivalence means that at a certain level, one establishes a formal correspondence between source-language and target-language units, without first making assumptions on the origins or the nature of the correspondence. This can be considered as a performance-based equivalence model, in comparison to translation studies' traditional, competence-based theories of equivalence, namely literal, functional and syntagmatic equivalence, total translation (Catford 1965, Jakobson 1959), and formal or dynamic equivalence (Nida 1964). Of the existing theories, the closest one is the theory of equivalence levels (Komissarov 1990).

The obvious performance-based formal equivalence model is a parallel corpus. Utterances arranged into man-made parallel corpora are considered *a priori* equivalent by researchers of computational techniques. However, the direct use of such corpora provides only a primitive surface-level imitation of the translation process: the computer does not imitate the transfer operations of the hu-

man translator; it only substitutes target-language utterances for source-language ones. Consequently, the translation process remains a black box.

A new model of translation equivalence. Quality assurance of translation – i.e. the fact that translations are being corrected – clearly shows that, according to general agreement, one source-language text can have different target-language translations that are „more equivalent" or „less equivalent". A simplistic representation of this is the continuum of translation that places the result of text transformations (i.e. translation and correction phases) along a straight line. Movement from left to right on the line signifies the target-language text approximating the target-language norm.

Comparing the first stage of the target-language text and the published text results in a correction or proofreading corpus. This can be used as an equivalence model because it uncovers the process of corrections, and compares the mental image of equivalence as employed by the translator and the proofreader.

3.2. The New Environment of Translation: the Origin of Translation Technology

This section describes translation technology as one of the research subjects in translation studies. It assesses how the micro- and macro-strategies of translation technology affect the translation process, with special regard to translation technology processes that were established to alleviate the effects of the changes in the environment.

The environmental changes can be briefly described by observing that the number and size of translation tasks have increased, and, at the same time, the available time to complete a task has fallen dramatically. Consequently, it can be proved that the quality of translations available on the market has significantly declined: the existence and evolution of translation technology offers a possible remedy to this situation.

3.3. The Micro-Strategy of Translation

To assess the micro-strategy in translation technology, one must assume that the target-language text is created in a computational translation environment using a translation memory. This section deals with the interaction between translation memory use and translation studies, by seeking answers to the following questions:

(4) To what extent can the use of translation memories considered as a model of transfer operations as described in the theory of equivalence levels?

(5) How is the translation process affected by the use of translation memories?

(6) What are the disadvantages of translation memory use, and how can these disadvantages be alleviated?

1. In short, translation memory use cannot be considered as a model of transfer operations. Currently existing tools only record the input and the output of the transformation, and are capable of performing only one manipulative action, namely, substitution, at the level of terms and segments (in most cases, sentences).

As a detour, we can observe that certain methods of machine translation are far better models for transfer operations than those in computer-assisted translation. Some machine translation methods are directly applicable as productive models to certain aspects of transfer operations. With computer-assisted translation, however, the human translator is allowed to perform the actual transfers, while in the former case, this task – i.e. the decision on transfer operations – is assigned to an automaton. There is yet another crucial difference: in a way, tools of computer-assisted translation record and replay the transfer decisions, while machine translation systems perform a combination of transfer operations according to their own algorithmic decisions. This difference is the reason for the apparent poorer quality of the latter: the algorithms and models used therein are suitable for explaining why machine-translated text is a worse approximation of the target-language norm than human translation.

2. If we consider the use of translation memories as a model of transfer operations, computer-assisted human translation involves recording and replaying sequences of transfer operations. The computer offers one or more possible translations that were stored earlier with a recurring source-language segment. The task of the translator is correcting the segment because (1) the current source segment may differ from the stored one, and thus the stored translation differs from the desired one; (2) the current source segment occurred earlier in a different context; (3) the stored translation is incorrect. One can safely assume that this requires a different skill than „clean" translation because the translator may have to revoke one or more supposed transfer operations. This is a certain type of proofreading: however, the goodwill hypothesis, i.e. that the translation offered from the database was originally intended as the equivalent of the current source segment, is no longer valid.

3. There are two objectives of research in the further development of translation memory systems:

- Increasing efficiency: Improving the quality and increasing the rate of translation memory matches by researching and introducing new proximity search methods. Chapter 4 provides details on the computational methods of assessing and measuring the efficiency of translation memories.
- Post-editing: Development of algorithms that modify the target-language segment taken from the database in such a way as to transform it into the proper translation of the current source-language segment. It is also a priority in machine translation research (cf. Isabelle et al. 2007, Kranias et al. 2004, Hodász G. et al. 2004).

3.4 The Macro-Strategy of Translation Technology

The macro-strategy of translation establishes a systematic framework for arranging the operations of micro-strategy into a well-defined process. It has three purposes:

- Ensuring the accomplishment of the new type of translation tasks (those of larger volume and tighter deadline) by means of organization and workflow management, as well as the establishment and enforcement of a project budget;
- Ensuring the technical synergy of translation, i.e. the co-ordinated transfer of source and target-language documents and the co-ordinated use of networked translation resources. Complex translation projects include the technical preparation of the target-language text (e.g. the conversion from PDF files) and the final technical production of the target-language text (e.g. typesetting or implementation of target-language website);
- Quality assurance: alleviating the negative effects of the time pressure. In addition to post-translation checks, this can also involve quality assurance measures performed in the preparation and the execution phase (i.e. during translation). The most sophisticated quality assurance methods pertain to terminology management, and, as such, they are covered in detail in Chapter 5 (Section 5.1).

Section 3.4 outlines the components of the macro-strategy, and then proposes two methods that provide for proper quality assurance in translation projects performed in a parallel manner, or under time pressure. The two methods are simultaneous proofreading and the automation of proofreading.

4 Corpus Linguistics and Natural Language Processing in Translation Technology

4.1 General Observations

When assessing the connection with translation technology, I consider natural language processing (or human language technology) as an applied field of computational linguistics.

As research on computational linguistics was initially motivated by the objective to create machine translation, it could as well have become closely related to translation technology. However, translation technology equipment makes little use of tools originating from natural language processing – neither in micro-, nor in macro-strategy –, although it is proposed in various forms by numerous authors (e.g. Hodász G. et al. 2004, Callison-Burch et al. 2005). Translation technology, in turn, employs several searching and text manipulation methods that are not related to natural language processing: they do not attempt to make use of, or manipulate the linguistic structure of the text.

If we look at the connection from the opposite direction, state-of-the art computational linguistics and language technology makes use of translation technology in a number of areas, precisely because existing translations provide a performance-based model to study translations, as opposed to the quasi-competence models currently used in translation studies. This is the same approach to translation as corpus linguistics has towards the general modelling of language. Translation performed under the surveillance of translation technology produces a large amount of parallel corpora, aligned at the segment level, which can be „mined" according to plenty of parameters. A possible interpretation of the ALPAC report (Pierce et al. 1966) and reports originating from XEROX (Kay 1980) leads to the suggestion that the efficiency of translation be enhanced by recycling performed human translations, instead of employing speculative computational models. The evolution of large-volume parallel corpora facilitated the development of statistical machine translation (cf. Callison-Burch et al. 2004).

This general section of Chapter 4 provides details on how methods and tools of natural language processing are being used in translation technology. The section deals with parallel corpora, text alignment and the concordance functionality of translation memories.

4.2 The SZAK Proofreading Corpus

I am convinced that the best way to model the proofreading of translations is the analysis of performed corrections. To analyze the corrections, one must study the differences between the target-language text as delivered by the translator, and the target-language text that was accepted for publication. This requires a parallel corpus that includes the first version of the target-language text and the published target-language text.

Over the course of translation and research activities of the SZAK publishing company, such a corpus was constructed, consisting of the text of technical books and web sites in the field of computing. The corpus consists of approx. 1.3 million words per language. The corpus has three components: it consists of the source-language text, the first target-language text and the published text.

Section 4.2 starts with the description of the characteristics and the construction of the SZAK Proofreading Corpus. This is followed by details on the modified Levenshtein algorithm (Levenshtein 1965) devised to reconstruct the correction operations by comparing the two versions of the target-language text.

4.3 Evaluation of Translation Memories and Enhancement of their Efficiency

The purpose of the translation memory is the quick recognition of (partially) recurring segments in a translator's work or within a translation team. It should then offer the earlier translation of the source segment found. The translation memory is implemented as a database that stores earlier source-language texts

and their translations. The translation memory is being transparently built while the translator is writing the target text in the translation environment.

Translation memories arrange segments into translation units. A translation unit (TU) is a pair of one source-language segment (in most cases, one sentence) and one or more target-language equivalents.

Chapter 4 demonstrates that translation memories are suitable for speeding up the translation of technical texts because they are far more homogeneous than literary texts (i.e. they contain more repetitions).

Measuring the efficiency of translation memories. The efficiency of the translation process can be measured by evaluating the time required to translate a specific unit of text. We then assume that this time decreases when translation memories are introduced.

When using translation memories, however, we can also observe that only a subset of the source text needs to be translated from scratch. The measure of the efficiency of a translation memory is the ratio of the size of this subset and the entire source text. If we adapt the custom of corpus linguistics, and represent the size of text in words, this means the following (using my denotations):

$$\eta = 1 - \frac{w_f}{w_t}$$

where η is the measure of efficiency; w_f is the sum of the size of source segments to translate (in words); w_t is the total size of the source text (in words). Translation memory efficiency equals 0 (zero) if all words must be translated, and, for example, 0.2 if there are automatic suggestions for 20% of the source words – so only the remaining 80% needs to be translated. Translation memory vendors specify this number when they communicate the increase of efficiency that can be achieved by using their product.

Section 4.2 refines the efficiency model, and proposes a method for quantifying the efficiency increase originating from the use of translation memories.

Limitations of character-based translation memories. The publicly available translation memories are character-based: they usually evaluate the similarity of segments by comparing trigraphs and bigrams (triples of characters and doubles of words). This method is similar to the statistical methods of computational linguistics. Character-based translation memories (cf. Navarro 2001, Navarro et al. 2001, and Planas 2000) evaluate the similarity of full segments only. As a result, a large proportion of stored segments remains hidden during translation. The longer the current source segment, the lower the probability of finding an appropriately close match in the database.

However, the translation memory can contain segments that are identical or similar to a part of the current source segment, and there can also be segments whose vocabulary is different but their syntax is analogous.

The thesis proposes two solutions to this problem. One is publicly available at the time of writing, while the other is in an experimental phase:

- Fragment search on the basis of character sequences (an extension to the existing translation memory technology);
- Linguistic decomposition or the syntax-based translation memory.

The syntax-based translation memory. Based on former research I was involved in (see Hodász G. et al. 2005, Hodász G. et al. 2004), this section describes a translation memory management method where the similarity of segments is determined on the basis of their linguistic structure, and is capable of dynamically substitute individual structural units in the target segment. The method is evaluated in detail. Based on the evaluation, the syntax-based translation memory constitutes a significant improvement over the character-based translation memory because

(1) it is capable of offering a target segment of adequate content in situations when the contents of the suggestions of the character-based translation memories are very different from the current source segment – or the character-based translation memory cannot offer any suggestions at all;

(2) there is less demand for post-editing as the morphosyntactic properties of the noun phrases are adjusted to the structural skeleton, and the potential correction points are marked. In other words, the suggestion can be corrected without first reading and fully interpreting it. From a psycholinguistic aspect this means that the correction can be performed by means of „shallower" linguistic operations, i.e. by translating smaller parts and correcting marked grammatical errors.

Evaluation criteria and methods of translation memories. Computational linguistics assesses the quality of language models. On that basis, a system of criteria could be established to evaluate translation memories as well. There are three quality measures:

(1) recall (r_m): referring to the efficiency of the querying algorithm, this number shows the proportion of the „dead weight" in the translation memory, i.e. the number of translation units (segments) that were stored but never retrieved. The algorithm can be assigned a reliable number by performing a measurement on multiple translation memory databases, and over a longer period.

(2) efficiency (η): the increase of translation efficiency as described earlier, given a specific translation memory database, a querying algorithm and a source text. Evaluated by measuring the actual effort of performing the translation. This number assesses the usefulness of computer assistance in translation.

(3) informativity (ι): evaluation of the numeric scores assigned to each suggestion by the translation memory, by determining how the score is related to the effort needed to correct the suggested target segment. This number assesses the quality of the query algorithm applied to the translation memory.

The thesis proposes methods for calculating and measuring the above numbers.

5 Translation Technology, Terminology and Lexicography

5.1 The Terminological Processes of Translation

The use of terminology is an integral part of translation. Translators and translation organizations collect words and expressions, build terminology databases, and occasionally publish dictionaries.

However, the literature of terminology studies rarely deals with the terminology problems of translation, although translation scholars often deal with terminological problems of various subject fields. In relation to translation, Sager (1990) uses the concept of secondary term formation that „[...] happens [...] as a result of knowledge transfer to another linguistic community which is carried out by means of term creation." (Sager 1990:80) Arntz (1993) recognizes that terminology research in translation might be problematic: „A detailed study of an individual phenomenon is often necessary in order to solve an acute translation problem. Investigations of this kind will frequently mention the neighbouring concepts without going into more detail, so that only a part of the field or system of concepts is handled." (Arntz 1993)

In general, we can observe that current descriptive terminology studies focus on the (structural) linguistic properties of terms, with the purpose of finding a linguistic model for the behaviour of terms, which can then be used in various applications such as automatic term extraction. Terminology-related sociolinguistic and onomasiological activities are rather of the prescriptive kind. A typical example is Pavel's article (Pavel 1993) that deals with neologisms and phraseology related to term formation, but, instead of the descriptive study of term formation processes, best practices are listed in connection with language teaching: an entire chapter is included on the linguistic conditions of the correctness of new terms. It can be observed that researchers involved in terminology are mostly nomenclators or meta-nomenclators, working on methods to provide precise descriptions and denominations for concepts, and to build conceptual systems.

As opposed to this, the thesis attempts at a descriptive approach to the principles and methods of term formation in translation.

The terminology processes of translation are described according to the following:

(1) The process of term formation in general: three different sources of primary and secondary term formation are described, including research and legislation; standardisation; and translation;

(2) The problem of term formation in translation: the concept of translation terminology is introduced (Lengyel 2006), and the requirement of term consistency is demonstrated;

(3) Possible strategies of term formation in translation: the continuum of possible strategies is described and aligned with the translation process. Accord-

ing to this, term formation can be performed at the preparation phase (total preparation), or at the post-processing phase (total proofreading), and quality assurance measures can be taken in the execution phase (supervised cooperation). Section 5.2 describes each strategy and the means of operating them;

(4) Possible technical means of terminology management in translation, ranging from terminology databases (term bases) to term extraction tools and query management systems.

5.2 Automatic Term Extraction

Term extraction has extensive international literature (e.g. Jacquemin 2001), and many applications. Section 5.2 provides an account of a development project and experiment (Kis, B. 2005) aiming at the development of a term extraction tool that is suitable for large translation tasks, and that can

- map the entire terminological skeleton of the source-language text, and
- shorten the time required to perform terminology preparation for a translation task.

A more precise formulation of the latter is that time required for text-specific terminology preparation must be substantially shorter than the time it takes for a human reader to read through the source-language text.

It is very important to note that the resulting term extraction technology is the first in supporting the Hungarian language. It also features a flexible linguistic framework that provides easy addition of further languages.

Section 5.2 describes the necessary corpus linguistics research, and demonstrates that the tool created meets both requirements. The description starts with the problem of defining terminology, especially from the aspect of computational models.

The most important feature of the resulting term extraction procedure and tool is that it employs more than one – statistical and rule-based – algorithms together, and provides a term candidate list where the items are assigned scores. In the list, the candidate terms can be categorized by score value ranges, and the categories can be used to predict the precision of the candidates falling into each category. Two experiments are described: the second one verifies the hypothesis of the categories. It was observed that there exists a category where the precision of the candidates is near 90%, which is far greater than the 30-50% usually attained by term extraction systems or raw frequency lists.

Irodalomjegyzék

Alonso, Juan Alberto. 2005. Machine Translation for Catalan-Spanish: The real case for productive MT. In: Hutchins-Kis B.-Prószéky (eds.): *Practical Applications of Machine Translation. Proceedings of the 10th EAMT Conference.* Budapest: Pázmány Péter Catholic University. 23-26.

Arevalillo, Juan José D. 2005. The EN-15038 European Quality Standard for Translation Services: What's Behind It? In: The Localization Insider 2005(04). Romainmôtier: LISA. http://www.lisa.org/globalizationinsider/2005/04/the_en15038_eur.html

Arntz, Reiner. 1993. Terminological Equivalence and Translation. In: Helmi B. Sonneveld, Kurt L. Loening (eds.): *Terminology. Applications in interdisciplinary communication.* Amsterdam-Philadelphia: John Benjamins. 5-19.

Austermühl, Frank. 2001. *Electronic Tools for Translators.* Manchester: St. Jerome.

Baker, Mona, Francis, Gill & Tognini-Bonelli, Elena (eds.). 1993. Text and Technology: in Honour of John Sinclair. Amsterdam-Philadelphia: John Benjamins. 233-250.

Berman, A. 1985. Translation and the Trials of the Foreign. Translation: L. Venuti. In: Venuti, L. (ed.) *Translation Studies Reader.* London: Routledge. 284-298.

Blank, Ingeborg. 2000. Terminology extraction from parallel technical texts. In: Véronis (ed.): *Parallel Text Processing – Alignment and Use of Translation Corpora.* Dordrecht-Boston-London: Kluwer Academic Publishers. 237–274.

Brown, Peter F., Della Pietra, Stephen A., Della Pietra, Vincent J., Mercer, Robert L. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation. In: Susan Armstrong (ed.): Using Large Corpora. Cambridge, Massachusetts: The MIT Press. 223-272.

Callison-Burch, C., Bannard, C., Schroeder, J. 2004. Improving statistical translation through editing. In: *Proceedings of the 9th EAMT Workshop.* Valletta: University of Malta. 26-32.

Callison-Burch, C., Bannard, C., Schroeder, J. 2005. A compact data structure for searchable translation memories. In: Hutchins, Kis, Prószéky (eds.): *Practical Applications of Machine Translation. Proceedings of the 10th Annual Conference of the European Association for Machine Translation.* Budapest: Pázmány Péter Catholic University. 59-65.

Carl, M. 2001. Inducing Translation Grammars from Bracketed Alignments. In: *Proceedings of the EAMT Workshop on Example-Based Machine Translation*. Elektronikusan elérhető: http://www.eamt.org/summitVIII/papers/carl.pdf

Castellví, M. T. C., Bagot, R. E., Palatresi, J. 2001. Automatic term detection: A review of current systems. In: Bourigault, D., Jacquemin, C. and L'Homme, M.-C. (eds.): *Recent Advances in Computational Terminology*. Amsterdam–Philadelphia: John Benjamins. 53–88.

Catford, J. C. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics.* 3. Oxford: Oxford University Press.

Choueka, Yaacov, Conley, Ehud S., Dagan, Ido. 2000. A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English. In: Jean Véronis (ed.): *Parallel Text processing. Alignment and Use of Translation Corpora.* Dordrecht-Boston-London: Kluwer Academic Publishers. 69-96.

Commission Of The European Communities. 2005 [EC 2005]. A New Framework Strategy for Multilingualism. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions. Brussels, 22.11.2005. COM (2005) 596 final. Elektronikusan elérhető: http://europa.eu.int/languages/servlets/Doc?id=913

Dróth Júlia. 2002. A fordítástudomány és a nyelvtudomány együttműködéséről. In: Fordítástudomány IV. 1. Budapest: Scholastica. 5-14.

Drugan, Joanna. 2004. Training Tomorrow's Translators. In: *Proceedings of the IV Conference on Training and Career Development in Translation and Interpreting*. Madrid: Universidad Europea de Madrid. http://www.leeds.ac.uk/cts/research/publications/leeds-cts-2004-02-drugan.pdf

Eco, Umberto. 1998(1993). *A tökéletes nyelv keresése.* Gál Judit, Kelemen János ford. Budapest: Atlantisz.

Eco, Umberto. 2001. *Experiences in Translation.* Toronto: University of Toronto Press.

Esselink, Bert. 2000. *A Practical Guide to Localization*. Amsterdam-Philadelphia: John Benjamins.

European Commission. 2004. Promoting language learning and linguistic diversity An action plan 2004-06. Luxembourg: Office for Official Publications of the European Communities.

Fillmore, Charles J. 1976. Frame semantics and the nature of language. In: *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280: 20-32.

Friedl, Jeffrey E. F. 2003. *Reguláris kifejezések mesterfokon.* Budapest: Kossuth Kiadó.

Fulford, H., Granell-Zafra, J. 2004. The freelance translator's workstation: an empirical investigation. In: *Proceedings of the Ninth EAMT Workshop.* Valletta: Foundation for International Studies, University of Malta. 53-61.

Gale, William A., Church, Kenneth W. 1993. A program for aligning sentences in bilingual corpora. In: Susan Armstrong (ed.): *Using Large Corpora.* Cambridge, Massachusetts: The MIT Press. 75-102.

Gerloff, P. 1987. Identifying the Unit of Analysis in Translation. In: Færch-Kasper (eds.) *Introspection in Second Language Research.* Clevedon: Multilingual Matters. 135-158.

Gibbon, Dafydd. 2005. *How to Make a Dictionary – Class notes.* Bielefeld University. Elektronikusan elérhető: http://wwwhomes.uni-bielefeld.de/~gibbon/Classes/Classes2005WS/HTMD/classnotes.html

Görz, G., Kesseler, M., Spilker, J., Weber, H. 1996. Research on Architectures for Integrated Speech/Language Systems in Verbmobil. Proceedings of COLING-96. Copenhagen.

Grőbler, T., Hodász, G., Kis, B. 2004. MetaMorpho TM: A Rule-Based Translation Corpus. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon.

Harris, S., Ross, J. 2006(2005). *Kezdőkönyv az algoritmusokról.* Bicske: SZAK Kiadó.

Haugen, Einar. 1983(1998). The implementation of corpus planning: Theory and practice. In J. Cobarrubias és J. A. Fishman (szerk.), Progress in Language Planning. Berlin: Mouton. 269–289. (Magyarul: A korpusztervezés kivitelezése: elmélet és gyakorlat. In Tolcsvai Nagy G., szerk. 1998: 143–160.)

Heltai Pál. 1999. Minimális fordítás. In: *Fordítástudomány I/2.* Budapest: Scholastica. 22-32.

Hodász, G., Grőbler, T., Kis, B. 2004. Translation memory as a robust example-based translation system. In: *Proceedings of the Ninth EAMT Workshop*. Valletta: University of Malta. 82-89.

Hofstadter, Douglas R. 1998(1980). Gödel, Escher, Bach. *Egybefont gondolatok birodalma.* Lipovszki Gábor ford. Budapest: Typotex.

Holmes, James S. 1988(1972). The Name and Nature of Translation Studies. In: James S. Holmes: *Translated! Papers on Literary Translation and Translation Studies*, Amsterdam: Rodopi. 67–80.

Homola, Petr, Kuboň, Vladislav. 2004. A Translation Model For Languages of Accessing Countries. In: Hutchins-Rosner (eds.:): *Broadening horizons of machine translation and its applications. Proceedings of the 9th EAMT Workshop.* Valletta: University of Malta. 90-97.

Horváth Ildikó. 2002. Nyelvi jogok és az Európai Unió nyelvpolitikája. *Fordítástudomány*, 2002. IV. évfolyam, 1. szám. Budapest: Scholastica. 15-47.

Hutchins, John. 1996. ALPAC: the (in)famous report. MT News International 14 (June 1996), 9-12. Reprinted in: S. Nirenburg, H. Somers and Y. Wilks (eds.:) *Readings in machine translation* Cambridge, Mass.: The MIT Press. 131-135.

Hutchins, John. 2003. Machine translation: general overview. In: Mitkov, Ruslan (ed.): *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press. 501-511.

Ieva Zauberga. 2005. Handling Terminology in Translation. In: Károly-Fóris (eds.): *New Trends in Translation Studies – In Honour of Kinga Klaudy.* Budapest: Akadémiai. 107-116.

Az IFIP ICC Információfeldolgozási értelmező szótára. 1968. Budapest: Országos Ügyvitelgépesítési Felügyelet. [IFIP-ICC 1968]

Isabelle, P. Goutte, C., Simard, M. 2007. Domain adaptation of MT systems through automatic post-editing. In: *Proceedings of the Electromagnetic Theory Symposium (2007) of the International URSI Commission B.* http://www.mt-archive.info/MTS-2007-Isabelle.pdf

Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge (Mass.): MIT Press.

Jakobson, R. 1959. On linguistic aspects of translation. In: R. Brower (ed.): *On Translation*. Cambridge (Mass.): Harvard University Press. 232-239.

John Hutchins. 2002. Machine translation today and tomorrow. In: Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz (eds.): *Computerlinguistik: was geht, was kommt? Computational linguistics: achievements and perspectives.* Festschrift für Winfried Lenders. Sankt Augustin: Gardez! Verlag. 159-162.

Jurafsky, D., Martin, J. H. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.

Kay, Martin. 1980. The Proper Place of Men and Machines in Language Translation. Xerox report CSL-80-11, Xerox Palo Alto Research Center.

Kay, Martin, Röscheisen, Martin. 1994. Text-translation Alignment. In: Susan Armstrong (ed.): *Using Large Corpora.* Cambridge, Massachusetts: The MIT Press.

Kilgarriff, A., Tugwell, D. 2001. Word sketch: extraction and display of significant collocations for lexicography. In: *Proceedings of the 39th ACL and 10th EACL workshop* 'Collocation: computational extraction, analysis and explotation'. Toulouse. 32–38.

Kis, Ádám, Kis, Balázs 2003. A prescriptive corpus-based technical dictionary. development of a multi-purpose technical dictionary. In: Pajzs, J. (ed.): *Papers in Computational Lexicography: Proceedings of COMPLEX 2003* Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences. 47–56.

Kis Ádám, Kis Balázs, Pohl Gábor. 2004. A számítógépes terminológiakivonatolás új megközelítése. In: *A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete.* Szeged: Szegedi Tudományegyetem. 63-72.

Kis Ádám, Kis Balázs. 2004. A szupermorféma. Nyelvtechnológia és szöveg. *In: A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete.* Szeged: Szegedi Tudományegyetem. 2004. 246-256.

Kis Ádám, Kis Balázs. 2004. Nyelvi tervezés a magyar informatikában. In: Balázs Géza (szerk.): *A magyar nyelvi kultúra jelene és jövője.* II. Budapest: MTA Társadalomkutató Központ. 155-170.

Kis Ádám. 1997. Gépszerű helyesírás. Az akadémiai helyesírási szabályzat és a számítógép. In: VII. Országos Alkalmazott Nyelvészeti Konferencia. Budapest: Magyar Elektronikus Könyvtár. http://www.mek.iif.hu/porta/szint/tarsad/nyelvtud/gepscikk/gepscikk.mek

Kis Ádám. 2004. Gyakorlati terminológia. In: Dróth Júlia (szerk.): *Szaknyelv és szakfordítás. Tanulmányok a Szent István Egyetem Alkalmazott Nyelvészeti Tanszékének kutatásaiból és kutatási témáiról.* Gödöllő: Szent István Egyetem. 46-52.

Kis Balázs. 2005a. Automatikus terminológiakeresés számítógéppel – kísérlet. In: *Fordítástudomány* 13. (VII. 1.). Budapest: Scholastica. 84-97.

Kis Balázs (szerk.). 2005b. *Angol-magyar informatikai fordítói szótár. A SZAK Kiadó szótára*. Bicske: SZAK Kiadó.

Kis Balázs, Lengyel István. 2003. Új módszerek az emberi fordítás gépi támogatásában. In: *Az I. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete.* Szeged: Szegedi Tudományegyetem. 268-275.

Kis Balázs, Lengyel István. 2005. A fordítás számítógépes segédeszközeiről. In: Emlékkönyv Klaudy Kinga 60. születésnapjára. Bicske: SZAK Kiadó.53-60.

Kis Balázs, Lengyel István. 2005. Fordítás-előkészítés és csoportos fordítás. In: Fordítók és Tolmácsok 3. Őszi Konferenciája, Budapest: MFE. 46–57.

Kis Balázs, Naszódi Mátyás, Prószéky Gábor. 2003. Komplex (magyar) szintaktikai elemző mint beágyazott rendszer. In: *Az I. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete*. Szeged: Szegedi Tudományegyetem. 145-153.

Kis, B., Villada Moirón, B., Bíró, T., Bouma, G., Pohl, G., Ugray, G., Nerbonne, J. 2004a. Methods for the Extraction of Hungarian Multi-Word Lexemes. In: Decadt, B. (ed.): *Proceedings of CLIN-2003.* Antwerp: University of Antwerp.

Kis, B., Villada, B., Bouma, G., Bíró, T., Nerbonne, J., Ugray, G. and Pohl, G. 2004b. A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-word Lexemes. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Lisbon. Vol. V, 1677–1681.

Kis, Balázs. 2002. Training Seminar on Translation and Localisation. Universitat Rovira i Virigli, Tarragona, Spain, 10-11 May, 2002. In: *Across Languages and Cultures 3 (2) (2002)*. Budapest: Szent Jeromos Alapítvány.

Kis, Balázs. 2004. Technology in the Translation Class: Introducing CAT Tools to Hungarian Translation Students. In: *IV Jornadas sobre la Formación y Profesión del Traductor e Intérprete.* Madrid: Universidad Europea de Madrid. http://www.uem.es/web/fil/invest/publicaciones/web/en/autores/ kis_art.htm.

Klaudy Kinga. 2006. *Bevezetés a fordítás elméletébe.* Budapest: Scholastica.

Klaudy Kinga. 1999. *Bevezetés a fordítás gyakorlatába.* Budapest: Scholastica.

Knuth, Donald E. 1994(1973). *A számítógép-programozás művészete. III. kötet: Keresés és rendezés.* Budapest: Műszaki.

Kranias, L., Samiotou, A. 2004. Automatic Translation Memory Fuzzy Match Post-Editing: A Step beyond Traditional TM/MT Integration. In: *Proceedings of LREC-2004.* http://www.mt-archive.info/LREC-2004-Kranias.pdf

Lengyel I., Kis B., Ugray G. 2004. MemoQ – Új megközelítés a fordítástámogatásban. Infrastruktúratanulmány. In: *A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete*. Szeged: Szegedi Tudományegyetem. 100–107.

Lengyel István. 2006. A nyelvi közvetítés szabványai – és ahogyan alkalmazzák őket. *In: Szent Jeromos-napi találkozások. Fordítók és Tolmácsok Őszi Konferenciája.* Budapest: MFE. 57-63.

Lengyel, István. 2006. Controlling the Workflow in Translation Projects. In: *MultiLingual Magazine, December 2006.* Sandpoint: MultiLingual Computing, Inc.

Левенштейн, В. И. 1965. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР 163.4:845–848. Appeared in English as: V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (1966):707–710.

Lewis, James P. 2005. *Project Planning, Scheduling & Control, Fourth Edition.* New York: McGrawHill.

LISA. 2007. TMX – Translation Memory Exchange. Version 2.0. Elektronikusan elérhető: http://www.lisa.org/standards/tmx/

LISA/OSCAR. 2007. Segmentation Rules eXchange (SRX) Version 2.0. Elektronikusan elérhető: http://www.lisa.org/standards/srx/

Matusov, E., Kanthak, S., Ney, H. 2005. Efficient statistical machine translation with constrained reordering. In: Hutchins, Kis, Prószéky (eds.): *Practical Applications of Machine Translation. Proceedings of the 10th Annual Conference of the European Association for Machine Translation.* Budapest: Pázmány Péter Catholic University. 181-188.

McConnell, Steve. 1996. *Rapid Development. Taming Wild Software Schedules.* Redmond: Microsoft Press.

McTait, K. 2001. Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns. In: *Proceedings of the Workshop on Example-Based Machine Translation.* http://www.eamt.org/summit_VIII/workshop-papers.html

Melby, Alan K. 1982. Multi-level translation aids in a distributed system. In: J. Horecký (ed.): *Proceedings of COLING 82.* Amsterdam: North Holland Publishing Company.

Melby, Alan K. 1995. Why Can't a Computer Translate More Like a Person? *Barker Lecture.* Elektronikusan elérhető: http://www.ttt.org/theory/barker.html

Melby, Alan K. 2000. Sharing of translation memory databases derived from aligned parallel text. In: Jean Véronis (ed).: Parallel Text Processing. Alignment and use of translation corpora. Dordrecht-Boston-London: Kluwer Academic Publishers. 347-368.

Miháltz, M., Prószéky, G. 2004. Results and Evaluation of Hungarian Nominal WordNet v1.0. In: *Proceedings of the Second International WordNet Conference (GWC 2004)*, Brno. 175-180.

Navarro, G. 2001. A Guided Tour to Approximate String Matching. In: ACM Computing Surveys, 33(1):31-88.

Navarro, G., Baeza-Yates, R., Sutinen, E., Tarhio, J. 2001. Indexing Methods for Approximate String Matching. In: *IEEE Data Engineering Bulletin,* 24(4), 19--27, Special issue on Managing Text Natively and in DBMSs.

Nida, Eugene A., Taber, Charles. 1982. The Theory and Practice of Translation. Brill Academic Publishers; New Ed edition.

Pavel, Silvia. 1993. Neology and Phraseology as Terminology-in-the-Making. In: Helmi B. Sonneveld, Kurt L. Loening (eds.): *Terminology. Applications in interdisciplinary communication.* Amsterdam-Philadelphia: John Benjamins. 21-34.

Pedersen, T., Banerjee, S. 2003. The design, implementation and use of the ngram statistics package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.* Mexico City. 153–158.

Pierce, John R., Carroll John B. et al. 1966. Language and Machines: Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC. Elektronikusan elérhető: http://darwin.nap.edu/books/ARC000005/html.

Planas, E., Furuse, O. 2000. Multi-Level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation. In: Proceedings of COLING-2000. Saarbrücken. 621-627.

Pohl Gábor. 2004. Iteratív bekezdés- és mondatszinkronizáció. In: Alexin Zoltán; Csendes Dóra (szerk.): *A II. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* Szeged: Szegedi Tudományegyetem. 117–123.

Pohl, Gábor. 2006. English-Hungarian NP Alignment in MetaMorpho TM. In: *EAMT-2006: Proceedings of the 11th Annual Conference of the European Association for Machine Translation*. Oslo. 69-74.

Prószéky Gábor, Kis Balázs. 2004. A nyelv és a számítógép. In: Kenesei I. (szerk.): A nyelv és a nyelvek. Budapest: Akadémiai Kiadó. 171-189.

Prószéky, G., Kis, B. 2002. Context-Sensitive Dictionaries. In: Shu-Chuan Tseng (ed.) *Proceedings of the 19th International Conference on Computational Linguistics* (COLING-2002), Vol. II, 1268-1272.

Prószéky Gábor, Kis Balázs. 1999. Számítógéppel emberi nyelven. Természetes nyelvi feladatok megoldása számítógéppel. Bicske: Szak Kiadó.

Prószéky, G., Tihanyi, L. 2002. MetaMorpho: A Pattern-Based Machine Translation Project. In: *Translating and the Computer 24.* London: ASLIB.

Prószéky, Gábor (1999): Language Technology Tools in the Translator's Practice. In: *Journal of Computing and Information Technology* Vol 7(3).

Prószéky, Gábor. 1996. Syntax As Meta-morphology. In: *Proceedings of COLING-96.* Copenhagen. Vol.2. 1123-1126.

Prószéky, Gábor. 2002. Translation of EU Documents. *Across Languages and Cultures* 3(2). Budapest: Szent Jeromos Alapítvány.

Pym, Anthony. 1993. Alternatives to Borders in Translation Theory. In: Susan Petrilli, (ed). 2003. *Translation Translation.* Amsterdam & New York: Rodopi. 451-463.

Rey, Alain. 1995. Linguistic and Terminological Standardisation from the Perspective of they Legal Status. In: Alain Rey: *Essays on Terminology* (ed. by Juan C. Sager). Amsterdam-Philadelphia: John Benjamins. 167-179.

Rirdance, Signe; Vasiljevs, Andrejs (eds.). 2006. *Towards consolidation of European terminology resources.* Experience and recommendations from EuroTermBank Project. Tilde: Riga. (TermNet Books)

Sager, Juan C. 1990. *A Practical Course in Terminology Processing.* Amsterdam-Philadelphia: John Benjamins.

Sager, Juan C. 1994. Language Engineering and Translation. Consequences of automation. Amsterdam: John Benjamins.

Schütz, Jörg. 1995. Terminological Knowledge in Multilingual Language Processing. European Commission. 1.–3. 1-66.

Sinclair, J., Hoelter, M., Peters, C. (eds). 1995. The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing. Studies in Machine Translation and Natural Language Processing. European Commission.

Steiner, George. 2005(1978). Bábel után. Nyelv és fordítás. 1. kötet. Bart István ford. Budapest: Corvina.

Szabari Krisztina. 1996. Az Európai Unió és a nyelvek. A nyelvi szabályozási gyakorlat, valamint a fordítás és tolmácsolás jelene és jövője. In: *Modern Nyelvoktatás.* 2(3). Budapest: Corvina. 31-45.

Szépe György. 2001. Nyelvpolitika: múlt és jövő. Pécs: Iskolakultúra.

Takeda, Koichi. 1996. Pattern-Based Context-Free Grammars for Machine Translation. In: *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.* Santa Cruz. 144-151.

Turcato, D., Popowich F. 2001. What is Example-Based MT? In: *Proceedings of the Workshop on Example-Based Machine Translation.* http://www.eamt.org/summitVIII/workshop-papers.html

Várnai Judit, Számadó Tamás (szerk.) (2004): Az Európai Unió hivatalos kifejezéstára. Bicske-Budapest: SZAK Kiadó-MorphoLogic.

Véronis, Jean (ed.) 1998. ARCADE sentence track data. In: *ARCADE-ROMANS-EVAL. Data from the 1998 evaluation exercise.* http://sites.univ-provence.fr/~veronis/data/arcroman98/Documentation/Introduction.htm

Véronis, Jean, Langlais, Philippe. 2000. Evaluation of parallel text alignment systems. The ARCADE project. In: Jean Véronis (ed.): *Parallel text processing.* Dordrecht-Boston-London: Kluwer Academic Publishers. 369-388.

Wagner, A. R., Fischer M. 1974. The String-to-string Correction Problem. In: *Journal of the ACM*, Vol. 21, #1. 168-173.

Wüster, Eugen. 1979. *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie.* Wien/New York, vol. I–II.

Jegyzetek

[1] technology: the application of scientific knowledge for practical purposes, especially in industry; [...] machinery and equipment developed from such scientific knowledge; the branch of knowledge dealing with engineering or applied sciences. [New Oxford Dictionary of English, 2001]

technológia: [...] a gyártási folyamat elmélete és gyakorlata. [...] [ÉKSZ 2003]

Технология (от греч. téchne — искусство, мастерство, умение и греч. логия — изучение) — совокупность методов и инструментов для достижения желаемого результата; способ преобразования данного в необходимое. [technológia: módszerek és eszközök halmaza a kívánt eredmény eléréséhez; módszer a meglevőnek a kívánttá való átalakítására] (Wikipedia: http://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)

[2] A „terminológia" szó a „-lógia" utótagban elvileg már utal a „tan"-ra, ám a terminológia közkeletű értelmezése nem ez: adott szakterület terminusainak halmazát jelenti. Ezért fontosnak tartottam, hogy a kutatási területet megkülönböztessem a nyelvi eszköztől, így került a megnevezés végére a „-tan" utótag. Az ezzel kapcsolatos szóhasználatot részletesen az 5. fejezet tisztázza.

[3] „Korábban már megjegyeztük, hogy miközben általános tudományos szövegek fordításához rendelkezésre áll géppel támogatott fordítás, nem rendelkezünk használható gépi fordítással. Emellett ennek közvetlen vagy megjósolható perspektívája sem látható." (Saját fordításom)

[4] „A gépi fordítás iránti érdeklődést ezután évekig illendő volt titokban tartani; éppen csak szégyellni nem kellett. A gépi fordítás «bukását» sokan még napjainkban is vitathatatlan tényként kezelik." (Saját fordításom)

[5] „[...] egyetérthetünk azzal, hogy az ALPAC-nak igaza volt, ami a gépi fordítással kapcsolatos szkepszisét illeti: a minőség kétségkívül rossz volt, és úgy tűnt, nem igazolja a kapott pénzügyi támogatás mértékét." (Saját fordításom)

[6] „Az ALPAC [...] hibáztatható azért, hogy kizárólag az amerikai tudomány és az amerikai kormányzat fordítási szükségleteit tartotta szem előtt, és nem ismerte fel az ipar és a kereskedelem szélesebb körű igényeit a már akkor is globalizálódó gazdaságban." (Saját fordításom)

[7] „A gépi fordítás egyre növekvő közvetlen felhasználása megmutatja, hogy az ilyen típusú alapvetően mechanikus működés nem válthatja ki a fordító ember gondolati folyamatait, ezzel is hangsúlyozva a fordítás minőségének fontosságát." (Saját fordításom)

[8] „Másfelől a professzionális fordító emberek sokféle szövegről tudnak jó fordítást készíteni. Az emberek sokféle szöveget tudnak kezelni; a számítógépek nem." (Saját fordításom)

[9] „A helyzet az, hogy a gépi fordítás olyan probléma, amely még egyáltalán nincs megoldva" [...] „Az emberi nyelvek jelenlegi elméleteiből [...] hiányzik egy létfontosságú tényező" (Saját fordításom)

[10] Az 1.1. és az 1.2. ábra a kibocsátó és a befogadó rendszerelméleti megnevezéseit (forrás és nyelő) is tartalmazza.

[11] „Az emberi nyelvek jelenlegi elméleteiből hiányzik egy létfontosságú tényező.[...] Ez a létfontosságú tényező a cselekvőképesség. A cselekvőképesség alatt azt a képességet értjük, amely lehetővé teszi, hogy az akaratunk által valódi döntéseket hozzunk, ezen belül olyan etikai döntéseket, amelyekért felelősek vagyunk." (Saját fordításom)

[12] „A rossz emberi fordítás azért érdekes, mert bár ember készítette őket, a gép által végzett fordításhoz hasonló módon." (Saját fordításom)

[13] Ismertetés a British Museum honlapján: http://www.britishmuseum.org/explore/highlights/highlight_objects/aes/t/the_rosetta_stone.aspx

[14] „A nyelvi diverzitás az Európai Unió egyik meghatározó jellemzője. Az Unió nyelveinek diverzitása iránti tisztelet az Európai Unió alapelve." (Saját fordításom)

[15] Románia és Bulgária 2007. évi belépése után.

[16] „A Bizottság arra a következtetésre jutott, hogy jelentős erőfeszítéseket kell tenni [...] annak biztosítására, hogy az anyanyelvén kívül mindenki még legalább két nyelvet beszéljen [...]" (A szerző fordítása)

[17] Az itt leírtak a saját műhely(ek) tapasztalatait tükrözik.

[18] Idézet a MATT honlapjáról: http://www.matt.hu/index2.htm

[19] Ahogy Hutchins (2002) írja: „More powerful PCs have encouraged the marketing of translation software for the general public. As general-purpose systems, the quality is inevitably poor. Input texts often contain high proportions of non-technical, colloquial language of the kind which MT systems have always found most problematic." („A nagyobb teljesítményű személyi számítógépek a nagyközönség számára is hozzáférhetővé tették a gépifordító-programokat. Ha általános célú rendszerként használják őket, a minőségük elkerülhetetlenül rossz lesz. A bemeneti szövegek jelentős mennyiségben tartalmaznak nem szakmai, köznyelvi elemeket, amelyek a gépifordító-rendszerek legnagyobb problémáját jelentik." – Saját fordításom)

[20] APSIC Comparator (http://www.apsic.com/en/products_comparator.html), Yamagata QA Distiller (http://www.qa-distiller.com/), MemoQ QA (http://www.kilgray.com/kilgray/companies/memoq?locale=hu)

[21] Az Assocation for Computational Linguistics (ACL) szerint: minden kutatási tevékenység, amely a nyelvészet és a számítógép-tudomány területeit érintő problémákkal foglalkozik. (Az ACL önmeghatározása szerint: „[...] international scientific and professional society for people working on problems involving natural language and computation")

A Wikipedia szerint: „Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. This modeling is not limited to any particular field of linguistics." (A számítógépes nyelvészet olyan interdiszciplína, amely természetes nyelvek számítógépen történő, statisztikai vagy szabályalapú modellezésével foglalkozik. Ez a modellezés nem korlátozódik a nyelvtudomány egyik konkrét területére sem. – *Saját fordításom)* (http://en.wikipedia.org/wiki/Computational_linguistics)

Dafydd Gibbon (2005) szerint: „[...] the interdisciplinary field which involves both linguistics and computer science, and is concerned with 1. automatising the analysis of text and speech corpora, 2. developing precise models of grammars and lexica which can be processed automatically." (Az az interdiszciplína, amely a nyelvészetre és a számítógép-tudományra épül, kutatásának tárgya pedig 1. az írott és a beszélt nyelvi korpuszok analízise, illetve 2. precíz, nyelvtanokból és lexikából álló modellek kialakítása, amelyek lehetővé teszik a [szövegek] automatikus feldolgozását. – *Saját fordításom)* (http://wwwhomes.uni-bielefeld.de/~gibbon/Classes/Classes2005WS/HTMD/classnotes.html)

[22] Amerikai székhelyű, kimondottan fordítással és lokalizációval foglalkozó tanácsadó szervezet. Webcíme: http://www.commonsenseadvisory.com/

[23] A LISA (Localization Industry Standards Association) nevű szervezet kidolgozta az SRX (Segmentation Rule eXchange) nevű szabványt, amely lehetővé teszi, hogy különböző gyártók

számítógépes programjai átadhassák egymásnak a mondatszegmentálási szabályokat. A szabvány szövege az értekezés írásakor itt érhető el: http://www.lisa.org/standards/srx/.

[24] További tudnivalók például a Wikipedián találhatók: http://en.wikipedia.org/wiki/Keylogger

[25] TMX: Translation Memory Exchange, a LISA (Localisation Industry Standards Association) nevű szervezet által létrehozott szabvány. Webcíme: http://www.lisa.org/standards/tmx/

[26] http://www.multicorpora.com/products/multiTrans4/

[27] http://www.apsic.com/en/products_xbench.html

[28] Adrian Kingsley-Hughes, Kathie Kingsley-Hughes. 2006. Beginning Programming. Indianapolis: Wiley Publishing.

[29] William R. Stanek. 2008. Microsoft Windows Server 2008 Administrator's Pocket Consultant. Redmond: Microsoft Press.

[30] „Secondary term formation [...] happens [...] as a result of knowledge transfer to another linguistic community which is carried out by means of term creation." (Sager 1990:80) (Saját fordításom)

[31] „A detailed study of an individual phenomenon is often necessary in order to solve an acute translation problem. Investigations of this kind will frequently mention the neighbouring concepts without going into more detail, so that only a part of the field or system of concepts is handled." (Arntz 1993) (Saját fordításom)

[32] „Standardisation is a separate process and consists of users reaching 'public' agreement to adopt a given term for use in specific circumstances." Sager (1990:114).

[33] „A trade-off triangle with schedule, cost and quality at its corners is a general management fundamental." (McConnell 1996:126) (Saját fordításom)

[34] A legteljesebb ilyen vita a Windows Vista operációs rendszer honosításakor, 2006 őszén-nyarán zajlott. A viták ugyan nem nyilvánosak, az anyag ugyanakkor kitűnően felhasználható a terminusalkotás elemzéséhez.

[35] Az értekezés írásakor a rendszer elérhető a http://www.eurotermbank.com/ címen. A projektum dokumentumai a http://project.eurotermbank.com/DesktopDefault.aspx weblapon találhatók.

[36] Ez a fejezet a Fordítástudomány 7. (2005) 1. számában megjelent cikk jelentősen kiegészített, javított változata.

[37] „In the onomasiological approach and in the search for translation equivalents, however, the choice between forms must be made via contextual examples which are translated into rules of usage." (Sager 1990:59, saját fordításom)

[38] A kísérlet részben az IKTA-00181/2003. számú, a Magyar Köztársaság Oktatási Minisztériuma által támogatott projekt keretében történt.

A kutatás jelentős inspirációt merített a 2004. december 17-én befejezett, 048.011.040. számú OTKA-NWO projektből. E projektet – holland-magyar bilaterális együttműködés keretében – a Groningeni Egyetem (Rijksuniversiteit Groningen) és a MorphoLogic valósította meg. A projekt témája többszavas lexémák szövegbeli keresése volt. (Lásd még: Kis B. et al. 2004a,b)

[39] New Oxford Dictionary of English, Oxford University Press, Oxford, 1998.

[40] A magyar nyelv értelmező szótára. Akadémiai, Budapest, 2003.

[41] Kis Balázs (2003): Windows XP haladókönyv. SZAK, Bicske.

[42] Prószéky Gábor-Kis Balázs (1999): Számítógéppel emberi nyelven. SZAK, Bicske.

[43] Adrian Kingsley Hughes-Kathie Kingsley-Hughes (2005): Beginning Programming. Wiley, Indianapolis.

[44] William R. Stanek (2003): Microsoft Exchange Server 2003 – A rendszergazda zsebkönyve. SZAK, Bicske.

[45] McConnell 1996 (lásd az Irodalomjegyzéket)