

Nyelveken és kultúrákon átívelő
többnyelvű információkeresés

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Bölcsészettudományi Kar

Könyvtár- és Információtudományi Intézet

Nyelveken és kultúrákon átívelő többnyelvű információkeresés

Témavezető:

Barátné dr. habil. Hajdu Ágnes
tanszékvezető egyetemi docens

Készítette:

Vig Dániel
Informatikus könyvtáros MA

A dokumentum lezárásának dátuma: 2015.01.04.

2015

Tartalomjegyzék

1. Bevezetés a többnyelvű információkeresésbe	5
1.1 A többnyelvű információkeresés problémái.....	6
1.2 Lekérdezés fordítása a dokumentum lefordítása ellenében	7
1.3 Közbeeső nyelv és közvetítő nyelv használata.....	8
1.4 A fordítás megközelítése többnyelvű információkeresés során	9
1.5 Szükséglet a többnyelvű információkeresésre és a többnyelvűség	11
1.6 A többnyelvű információkeresés története.....	12
2. Gépi fordítás.....	13
2.1 Szabály alapú gépi fordítás.....	15
2.2 Statisztikán alapuló gépi fordítás.....	16
3. Tudás alapú technikák	17
3.1 Tezaurusz használata	18
4. Szótár alapú fordítás	19
4.1 Alap megközelítés	20
4.2 Kifejezések súlyozása.....	21
4.3 A szótárak terjedelme	22
4.4 Fordítási bizonytalanság.....	22
4.5 A fordított szavak kiválasztása.....	23
5. Párhuzamos és összehasonlító gyűjteményeken alapuló fordítás.....	24
5.1 Párhuzamos gyűjtemények.....	25
5.2 Bekezdés és mondat egyeztetése	25
5.3 Összehasonlító gyűjtemények	27
6. További módszerek, amikkel javítani lehet a többnyelvű információkeresést.....	28
6.1 Elő- és utófordításos kibővítés	28
6.2 Összefüggéstelen összekapcsolás	29
6.3 Egyesítő fordítások.....	29
6.4 Átívelő fordítás.....	30

6.5 Az egynyelvű és többnyelvű kapcsolatok egyesítése	31
7. Többnyelvű információkeresés többnyelvű ontológiák felhasználásával, melyek a fordítás és a lekérdezés bővítésül szolgálnak.....	32
7.1 Az ontológia a gépi olvasású szótárral szemben	33
7.2 Értékelés.....	34
8. Összegzés	35
Irodalomjegyzék.....	36

1. Bevezetés a többnyelvű információkeresésbe

Az információ keresése mindennapjaink része, legyen az szórakozás vagy munka céljából való használat. A nyelveken átívelő információkereséssel mindennap szembesülünk. Célja, hogy olyan dokumentumokat találjon meg a felhasználó, amelyek nyelve nem azonos a kereső személy által beszélt nyelvvel. A legtöbb esetben az emberek saját nyelvükön szeretnének releváns információt találni, amely általában a lekérdezés nyelve is. Azonban egyre jobban körülvesznek minket az olyan információk, melyek nem az anyanyelvünkön íródtak. Azonban az internet-hozzáférés lehetőségeinek kiszélesedésével, egyre jobban körülvesznek minket az olyan információk, melyek nem az anyanyelvünkön íródtak. Az internet hatalmas mennyiségű információhoz nyújt hozzáférést (Ungváry és Vajda, 2002.), melyekhez különböző médiákon – szöveg, kép, videó, hang – és nyelveken keresztül lehet hozzájutni. Ha például szeretnénk hozzáférni egy orosz nyelvű dokumentumhoz, melyre német nyelven teszünk fel lekérdezéseket, azonnal felmerül a többnyelvű információkeresés problémája, melynek lényege, hogy olyan információkat keressünk, melyek más nyelven vannak, mint a lekérdezés nyelve (Nie, 2010.).

A többnyelvű információkeresés annyiban különbözik a hagyományos információkereséstől – nevezzük a saját nyelven történő keresésnek –, hogy a felmerülő nyelvi akadályokat valahogyan át kell hidalni, de akárcsak a hagyományos keresés, a többnyelvű keresés is több összetett technológiából épül fel (Uszkoreit, 1998.). A hagyományos keresés és a többnyelvű keresés alapvető céljai megegyeznek, hiszen mindkét esetben egy bizonyos dokumentumot szeretne elolvasni valaki, a különbség csakis annyi, hogy a hagyományos keresés az egyén saját nyelvén történik (Uszkoreit, 1998.). A többnyelvű információkereséshez azonban szükséges a hagyományos keresés problémáinak ismerete is (Hollink, Kamps, Monz és de Rijke, 2004.). A nyelvi nehézségeket megoldhatjuk azzal, hogy a lekérdezéseket lefordítjuk a célnyelvre, vagy pedig fordítva, azaz a dokumentumokat fordítjuk le a forrásnyelvre.

A nyelveken átívelő többnyelvű információkeresésnek leggyakrabban három fő megközelítési módja lehetséges:

- Szövegfordítás gépi fordítási technikák útján.
- Tudás alapú technikák.
- Gyűjtemény alapú technikák.

Mindegyik technikának meg van a maga előnye és hátránya. Helyzetfüggő, hogy mikor melyik módszert alkalmazzuk, melyik a legjobb megoldás (Peters, 1997.).

Az információkereső rendszerek lehetőséget biztosítanak a felhasználóknak arra, hogy olyan információkat szerezzenek be, amelyek relevánsak a megszerezni kívánt információra nézve, és ezt lekérdezés formájában tehetik meg (Talvensaari, 2008.). A többnyelvű információkereső rendszerek hasznosak lehetnek a különböző szintű nyelvtudást birtokló felhasználóknál, pl. valaki képes megérteni a dokumentum célnyelvét, viszont nem képes azt használni, így nem is tud lekérdezéseket feltenni. A rendszerek használata bármilyen formában megnyilvánulhat, lehet pl. egy tudományos kutatás, munkával összefüggő keresés vagy barkácsolással kapcsolatos tippek beszerzése. Az információkereső rendszereket viszonylag jól fel lehet osztani, a legcélszerűbb terminológiai megközelítés a következő két csoport:

- rendszerorientált rendszerek és
- felhasználó-orientált rendszerek.

A többnyelvű információkeresés fejlesztése napjaink fontos feladata, mert ha sikerülne megoldani, hogy különösebb anyagi és munkaerő ráfordítás nélkül jussunk hozzá idegen nyelvű dokumentumokhoz – ehhez elsősorban gépi fordítást kellene alkalmazni –, akkor gyakorlatilag bárki bármilyen publikált, az interneten elérhetővé vált információhoz hozzáférhetne (Uszkoreit, 1998.).

1.1 A többnyelvű információkeresés problémái

Az információkeresés egyik alapvető problémája – még egy adott nyelven belül is –, hogy egy szónak nemcsak több jelentése van, hanem azokat különböző módokon is lehet értelmezni (Nie, 2010.). A dokumentumok és a lekérdezések kifejezések által vannak jelen a folyamatban, melyek az előbbieken fordulnak elő, s amelyek eltérő jelentéssel bírhatnak, annak ellenére, hogy ugyan azt jelentik, hisz vegyük csak például a magyarban lévő piros és vörös szavakat. A szakirodalmi meghatározás alapján ez a két szó ugyan azt a színt jelöli, ám valószínűsíthető, hogy az emberek többségének a pirosról egy élénkebb és intenzívebb, míg a vörösről egy sötétebb és mélyebb szín jut az eszébe. Az ilyen problémák megnehezítik, hogy összepárosítsuk a releváns dokumentumokat a lekérdezés kifejezéseivel.

Ez a probléma még jobban előjön, amikor a dokumentumok maguk is több nyelven vannak egy gyűjteményben:

- Large floods in Central Europe in 2013.
- Schwere Hochwasser in Mitteleuropa im 2013.
- Nagy árvizek Közép-Európában 2013-ban.

Ez a példa jól mutatja, hogy megegyező információkat összehasonlító módon kell kezelni, még ha különböző nyelveken is vannak leírva. Az egyik fő feladat a többnyelvű információkeresésnél, hogy olyan eszközöket fejlesszünk ki, melyek összepárosítják a különböző nyelven lévő, megegyező kifejezéseket, de legalábbis a legközelebbi hasonló jelentésű szavakkal kapcsolják össze őket.

1.2 Lekérdezés fordítása a dokumentum lefordítása ellenében

A fordításnak három lehetséges módja van:

- a lekérdezés lefordítása,
- a dokumentum lefordítása és
- a két módszer egyidejű használata.

A lekérdezés fordításakor a lekérdezést alkotó szavakat és kifejezéseket fordítjuk le a célnyelvre, míg a dokumentumfordítás során vagy a célnyelvi, vagy a forrásnyelvi dokumentumot fordítjuk le a másik nyelvre. A lekérdezés lefordításánál a gond az, hogy ha túl általános kulcsszavakat adunk meg, akkor túl sok dokumentumot kapunk meg, amelyek egyáltalán nem is relevánsak a keresés szempontjából, míg a dokumentumfordítás nagy problémája, hogy idő- és erőforrás igényes a pontos fordítás létrehozása.

Általános nézet, hogy a legmegfelelőbb megközelítés a lekérdezés lefordítása, ezért a többnyelvű információkereső-rendszerek jórészt ezt a megoldást használják, leginkább a következő ok miatt: adott például egy lekérdezés, ahol a felhasználó szabadon választhat a nyelvek közül, melyre szeretné lefordítani a lekérdezését. Abban az esetben, amikor a felhasználó képes megérteni a lekérdezés lefordított kifejezéseit, akkor arra is képes, hogy kijavítsa a fordítást, mielőtt felhasználná azt a dokumentumok keresésére. Ezzel időt és energiát spórol meg, valamint nem kezd el hiábavalóan keresni a dokumentumok között, amelyek közt a rossz lekérdezés miatt nem lesz releváns találat. Ez a megközelítés sokkal rugalmasabb és sokkal nagyobb interaktivitást tesz lehetővé a felhasználóval (Nie, 2010.).

A lekérdezés-fordítás azonban gyakran beleütközik a pontatlanság problémájába, melyet csak még tovább növel a rövid lekérdezések relatíve kis kontextusa (Peters, 1997.). A hatékonyságot növelheti, ha a lekérdezés előtt feltérképezzük a dokumentumot, mert abban az

esetben, ha ezt elmulasztjuk, akkor a lekérdezés csakis akkor érhet el eredményt, ha annak elemei pontosan úgy szerepelnek a dokumentumban, ahogyan azok a keresőképben vannak (*Uszkoreit*, 1998.).

Ha ebből a szemszögből vizsgáljuk meg a kérdést, akkor a dokumentum lefordítása máris létjogosultságot nyer, valamint akár pontosabb fordításokhoz is juthatunk, hála a dokumentumban lévő kontextus gazdag környezetnek. A mai hatékony gépi fordító rendszerek is lehetővé teszik mindezt, azonban ezek a rendszerek sem teszik nyilvánvalóvá azt, hogy képesek teljesen kihasználni a rengeteg kontextus adta lehetőséget a dokumentumfordítás során (*Nie*, 2010.).

Több tanulmány is készült a két módszer összehasonlítására, de egyiknél sem jutottak arra az eredményre, hogy valamelyik módszer hatékonyabb lenne a másiknál. Ennek oka az lehet, hogy a dokumentumfordító rendszerek csak korlátozott mennyiségű kontextuális információt képesek kinyerni, míg a mondatokat egyenként fordítják. Viszont van egy nagyon erős érv a dokumentumfordítás mellett: „off-line” is lehet végezni, nem kell hozzá se áram, se számítógép, csak egy ember, aki jól ismeri a nyelvet (*Hiemstra és De Jong*, 1999.).

Egy további fontos aspektusa a dokumentumfordításnak, hogy előre meg kell határozni, hogy milyen nyelvre szeretnénk lefordítani, valamint hogy az összes lefordított dokumentumot el kell tárolni (*Nie*, 2010.). Egy valóban többnyelvű információkereső környezetben – az ideális rendszer eléréséhez – az embernek arra kellene törekednie, hogy minden egyes dokumentumot lefordítson valamennyi nyelvre. Viszont ez nem célravezető a tárhely-igény megnövekedése és a többlet dokumentumok létrejötte miatt. Mindazonáltal, ha egy dokumentumot egyszer lefordítottak a lekérdezés nyelvére, akkor onnantól bármely felhasználó viszonylag könnyedén el tudja olvasni. Más esetben a keresés után szükséges a fordítás ahhoz, hogy értelmezhető, felhasználható legyen a talált dokumentum.

A kísérletek alatt mutatkozó korlátozott előny miatt, a legtöbb mai kutatás és fejlesztés a lekérdezés-fordításra koncentrál mintsem a dokumentumfordításra, köszönhetően az előbbi nagyobb flexibilitásának (*Nie*, 2010.).

1.3 Közbeeső nyelv és közvetítő nyelv használata

Az erőforrások szűkössége miatt nem mindig várható el a direkt fordítás két nyelv között. Lehetséges viszont, hogy ezt ki lehet küszöbölni úgy, hogy egy harmadik nyelvet iktatunk közbe a fordítás folyamatába. Ezzel kapcsolatban két fajta megközelítés lehetséges:

- a dokumentumot és a lekérdezést lefordítjuk e közbeeső nyelvre,
- vagy a dokumentumot vagy a lekérdezést fordítjuk le erre a nyelvre, s majd csak ezután a célnyelvre.

Ha az első megközelítést használjuk, akkor a közvetítő nyelv esetéről is beszélünk (Nie, 2010.).

Némely kutató megpróbált közbeeső nyelvet használni, s az eredmények azt mutatják, hogy hasznos lehet olyankor, amikor a direkt fordítás nem lehetséges. Azonban ha lehetséges a direkt fordítás, akkor az nyilvánvalóan sokkal hatékonyabb a helyettesítő fordításnál.

1.4 A fordítás megközelítése többnyelvű információkeresés során

Az egyszerű fordítás problémái mellett a többnyelvű információkeresés legnagyobb gondja maga a fordítás. Fordítás a következő két esetben szükséges:

- Az első, amikor a lekérdezést a forrásnyelven teszik fel és a felhasználó a célnyelvben keres dokumentumokat. Ekkor a forrásnyelvi kifejezéseket le kell fordítani a célnyelvre, vagy fordítva, és így válik elérhetővé a keresett információ.
- A másik eset, amikor a dokumentumokat a célnyelvről le kell fordítani a forrásnyelvre, hogy a felhasználó elolvashassa azokat.

Bár a két eset hasonló, mégis teljesen más problémákba ütközik a megoldásuk során a szakember (Nie, 2010.). A második esetben említett probléma a klasszikus dokumentumfordítás feladata, amelyre az automatikus gépi fordítás a legjobb technika. De a gépi fordítás nem igazán alkalmazható az első esetre a következő okok miatt:

- Kevesebb szigorú szintaxis szükségeltetik. A fordítás célja nem az, hogy olvashatóvá váljon az ember számára a dokumentum vagy a lekérdezés, hanem hogy a rendszer (azaz a gép) képes legyen összekapcsolni a lekérdezéseket a dokumentummal, avagy fordítva. Ezért elég, ha a fordítás, csupán a keresőrendszerben használható, ami gyakran kulcsszavakra épül és csak ezeket kell lefordítanunk. Ez azt jelenti, hogy nem kell alávetnünk magunkat a célnyelv szigorú nyelvtanának, amikor fordítunk, hanem kizárólag a lefordítandó kulcsszavak a fontosak (Nie, 2010.).

- Nagyobb a bizonytalanság a lekérdezés-fordítás esetén, mivel a lekérdezések általában nem állnak két vagy három szónál többől. Emiatt nagy a pontatlanság, s az ebből eredő zavar, mivel nehéz megtalálni a megfelelő kifejezéseket. Ez a kiválasztási probléma nem csak a fordítás szempontjából van jelen, hanem az egyszerű egynyelvű keresés esetében

is, amikor a releváns dokumentumokat akarjuk kiválasztani a gyűjteményből (*Ballesteros és Croft, 1998.*).

– A kívánt kibővítés elérése a lekérdezésben. Mint már szó volt róla, a dokumentumfordítás célja az, hogy szeretnénk egy másik nyelven is elérni a forrásban található információt. Ilyenkor nincs a birtokunkban a szavak igazi jelentése, csak következtethetünk rájuk a kontextusból és több lehetséges jelöltünk akadhat a helyes megfejtésre. Ahhoz, hogy a lekérdezést összepárosíthassuk a dokumentumban lévő különböző alakú, de megegyező vagy hasonló jelentésű szavakkal, kíváncsi, hogy a lefordított lekérdezés az összes lehetséges jelentés alternatívájával rendelkezzen. Ez a lépés már az egyszerű egynyelvű információkeresés során is előjött, de a többnyelvű információkeresésnél szinte állandóan jelen van (*Nie, 2010.*).

– Kifejezések „súlyozása”. A hangsúly, a fogalmak egymáshoz való viszonya, és rendszerben való elhelyezése rendkívül lényegbevágó az információkeresésben. Ekkor minden kifejezéshez egy adott értéket adunk, hogy jelezzük mennyire is fontos az összepárosítás folyamatában (*Talvensaari, 2008.*). Többnyelvű keresés esetén az is lényeges, hogy a lefordított kifejezésekkel is hasonlóan járjunk el. Ha ezt meg tesszük, akkor nem csak a lekérdezésben szereplő kifejezés fontosságával leszünk tisztában, hanem a fordítás helyességével is. Ez a módszer egy-egy keresőrendszer különlegessége, nem része a gépi fordításnak.

A fentiekből is kitűnik, hogy a többnyelvű információkeresés nem csak egy egyszerű fordítási feladat, hanem egy olyan művelet, mely mélyen beépült a keresésbe (*Nie, 2010.*), és bár sok probléma megegyezik az általános fordítással, mégis sok speciális kérdés és feladat van, melyekkel különböző úton és módszerekkel kell megbirkózni. Ezekkel általában két féle módon próbálkoznak:

- szótár alapú fordítással, melyben egy kétnyelvű szótár segítségével választjuk ki a lehetséges fordításokat a forrásnyelvből vagy
- a párhuzamos gyűjteményeken alapuló fordítással, ahol mind a forrásnyelv és mind a célnyelv dokumentumai megvannak. Ennél a megoldásnál a két nyelv közti fordítási kapcsolatokat próbáljuk meg kihasználni – legyen az egyes szavaknál vagy magasabb szinten, pl. kifejezéseknél. Ezeket a kapcsolatokat fel lehet használni a lekérdezések és a dokumentumok lefordítására.

Észre kell vennünk, hogy a fentebb említett módok nem szigorú szabályok szerint határozzák meg a fordítás mikéntjét, például a klasszikus gépi fordítás is kihasználja a párhuzamos gyűjtemények adta lehetőségeket.

1.5 Szükséglet a többnyelvű információkeresésre és a többnyelvűség

Bár az egynyelvű keresés létjogosultságának nincs szüksége további bizonyítékokra – elég csak az interneten elérhető keresőmotorok széleskörű használatára gondolni –, nem mindenki számára lehet egyértelmű, hogy valaki olyan információt szeretne elérni, mely más nyelven található (Nie, 2010.). Amint említettük már, az információkeresés célja, hogy releváns információkhoz jussunk. A leírás formája általában kisebb fontossággal bír (már ha az olvasó beszéli az adott nyelvet, érti a használt kommunikációs formát, alakot, szimbólumokat, stb.), pl. lehet szöveg, kép, videó, hang, stb. Elvben bármilyen releváns információ lényeges a felhasználónak, viszont ezek csak akkor hasznosak a számára, ha képes megérteni azokat. Ez az az ok, amiért a mai információkereső-rendszerek egy nyelven biztosítanak lehetőséget a keresésre, valamint indokolja azt is, hogy miért keresnek az emberek a saját nyelvű dokumentumok után (Nie, 2010.). De hála az automatikus fordításban elért fejlődésnek, egyre több esély van megszerezni olyan információkat is, melyek nem a felhasználó nyelvén vannak. A fordítási eszközöktől függetlenül, arra is szükség van, hogy olyan dokumentumokat találhassunk meg, melyek nem kötődnek a nyelvhez. Ezekre példa:

- Az információ lehet olyan formában, melyet a felhasználó képes értelmezni, de arra másik nyelven kell keresni. Ilyen eset lehet a multimédiás keresés, mikor pl. egy képre keres rá az ember. Ekkor a nem a nyelv a fontos, hanem maga a keresett objektum, azaz maga a kép. Bár a multimédiás keresés nem kapcsolódik szorosan a többnyelvű kereséshez, mégis, ha a szövegből derül ki, hogy az információ releváns, akkor már használnunk kell azt (Nie, 2010.).

- A kívánt információ nem létezik a felhasználó nyelvén, csak egy másik nyelven. Ilyenkor mindenképpen le kell fordítani a lekérdezést vagy a dokumentumot (vagy mindkettőt), valamint vissza is kell fordítani azokat a forrásnyelvre (Nie, 2010.).

- Amikor egy dokumentum több nyelven is leírnak lényeges dolgokat.

- Amikor a felhasználó minden releváns információhoz szeretne hozzájutni, az elérés nyelvétől függetlenül. Erre egy tipikus példa a szabadalmak utáni keresés, vagy amikor egy nemzetközi cég kutatást végez, hogy milyen más vállalatok tevékenykednek ugyan abban az üzleti szektorban, mint ők.

- Azon emberek, akik több nyelvet is képesek megérteni, illetve beszélni. Jó részük képes megérteni a dokumentumot, de nem képesek a megfelelő lekérdezések előállítására a tanult nyelven/nyelveken, így nekik is nagy segítség a többnyelvű információkereső-

rendszerek használata. Ezek a megoldások segíthetnek abban is, hogy gyorsabban hozzanak létre lekérdezéseket egyszerre több nyelvben is.

A fent említett szituációk manapság hétköznapi esetnek számítanak, hisz a globalizáció hatása már a világ szinte minden országára kiterjedt, s ezt akár tekinthetjük a globalizáció egyik előnyének is (Nie, 2010.).

1.6 A többnyelvű információkeresés története

Nem lehet pontosan időben behatárolni, hogy mikor is kezdtek el a többnyelvű információkeresés problémájával foglalkozni (Uszkoreit, 1998.), de az első tanulmányok már az 1960-as évek vége felé megjelentek témakörben, ld. Salton (1970) munkái. Az igazi nagy fellendülést az 1990-es években tapasztalhattuk, az internet gyors előretörésével párhuzamosan (Peters, 1997.). Annak ellenére, hogy a végrehajtott keresések nagy része egy nyelven történt (és még ma is fennáll ez a helyzet), az igény, hogy más nyelveken is kereshessenek a dokumentumra, egyre nőtt. 1997-ben a többnyelvű információkeresés végre hivatalos formát is nyert, amikor megtartották a TREC-6 konferenciát (Text Retrieval Conference), amit az amerikai Nemzeti Szabványügyi és Technológiai Hivatal szervezett. Azóta csupán csak ez a hivatal több konferenciát is szervezett, és egyre több nyelvet is beemelt az 1997-ben indult projektekbe. Természetesen a Nemzeti Szabványügyi és Technológiai Hivatal munkásságán túl számos más fórum, intézmény, kutatócsoport foglalkozik azóta is a többnyelvű információkereséssel, ill. a kérdéskör egy-egy speciális területével.

Az európai nyelvekkel kapcsolatosan a kutatások 2000-ben kezdődtek a CLEF (Cross-Language Experiment Forum) keretein belül. Az első kísérletek során angol, francia, német és olasz dokumentumok között kerestek, angol, francia, német, olasz, holland, svéd, finn és spanyol nyelvű kifejezések használatával. Az évek során egyre több nyelvvel hajtottak végre kísérleteket. 2005-től fogva az internetes gyűjteményekben is alkalmazták a módszereket és végeztek kutatásokat (pl. az Európai Unió szervezeteinek weboldalaiban). 2007-ben a szervezet kiterjesztette a vizsgálódásait az indiai nyelvekre is.

A japán Nemzeti Informatikai Hivatal által szervezett NTCIR 1999-ben kezdődött. Ők elsősorban az ázsiai nyelvekre koncentráltak. Mellettük a kínaiak is végeztek kutatásokat a többnyelvű információkeresés témakörében, pl. a Pekingi Egyetem 2004 óta teszi közzé eredményeiket a kínai webbel kapcsolatban.

Mindezen kutatások sokat hoztattak a többnyelvű információkeresés fejlődéséhez. Kezdetben az egy nyelvű információkeresés hatékonysága (melynek „mértékegysége” az átlagos precizitás átlaga) jóval meghaladta a többnyelvűét – közel kétszer hatékonyabb volt. Mára azonban a többnyelvű keresés képes elérni az egynyelvű keresés pontosságának szintjét. Ez az eredmény igazolja, és teljesen nyilvánvalóvá teszi a fenti kutatások sikerességét.

Egy másik tény, amiért sikeresnek ítéltjük meg ezeket a kutatásokat, hogy rengeteg kereskedelmi cég kezdett termékeket kínálni keresőrendszerekkel kapcsolatban. A Yahoo! 2006-ban indította el a fordítóját, a Google 2007-ben.

Természetesen e rendszerek pontossága nagyban függ a témától és a használt nyelvektől, a megfogalmazások bonyolultságától, tagoltságától, stb., de az vitathatatlan, hogy nagyban megkönnyítik a felhasználók dolgát, a más nyelvű dokumentumok hozzáféréssel kapcsolatban – és ha hozzávesszük az olyan keresőket, mint a Google Scholar (mely tudományos munkák után keres), akkor még inkább igaz ez az állítás (Nie, 2010.).

2. Gépi fordítás

A gépi fordítást két fő kategóriára lehet osztani:

- a tradicionális szabály alapú és
- a statisztikai alapú.

Például: a Systran szabály alapú fordító, míg a Google Translate és a Language Weaver statisztikai alapú (Nie, 2010.). A szabály alapú rendszerek manuálisan megállapított szabályok és eszközökkel operálnak, melyek különböző típusai: lexikai, kifejezéses, szintaktikus, szemantikus, stb. rendszerek. A kifejezések és a fordításaikat el lehet tárolni szótárakban, akár csak a szimpla szavakat. Például a francia „pomme de terre”-t és annak magyar megfelelőjét a „burgonya”-t el is kellene tárolni, hogy a megfelelő fordítás később is gyorsan rendelkezésre állhasson. A nyelvtani és szintaktikai szabályok segítenek felismerni a forrásnyelv szintaktikai szerkezetét és segítenek létrehozni a célnyelvben a megegyező szerkezetet. A szemantikai szabályok segítenek kiválasztani a helyes fordítást, amikor a forrásnyelvben fordítási bizonytalanság áll fenn. A szemantikai szabályokat azonban a legnehezebb felépíteni és beépíteni a fordítási folyamatba. Egy nyelv teljes szemantikai modelljének a felállítása hatalmas mennyiségű szemantikai információt igényel, ami majdhogynem egyenlő lenne a világ tudásának lemodellezésével. Egyértelmű, hogy ezt elég nehéz, vagy még inkább képtelenség a gyakorlatban megvalósítani.

A statisztikai alapú fordítás statisztikai nyelv és fordító modelleken alapul, melyeket automatikusan nyernek ki nagy mennyiségű szövegből és fordításaikból – azaz legalább két dokumentumcsoport van. Ezek a kinyert elemek lehetnek szavak, azok ragozott változatai, kifejezések, stb., melyek mindkét nyelvben és a fordításukban is szerepelhetnek.

A gépi olvasású osztályozás hagyományos megközelítésekor, különböző típusú nyelvspecifikus szükségletnek kell megfelelni, hogy végrehajthassuk az olvashatósági szintek beosztását. Némely nyelvénél ez a feladat eléggé kiforrott. Például a Laesbarhed-index a Nyugat-Európai nyelvekben a következő formulát használja: $RL_D = \text{szavak/mondatok} + 100 \times \text{szavak, melyek hatnál több karakterből állnak/szavak száma}$ (Kirchner, Nuger és Zhang, 2009.).

A D az a dokumentum, amely számunkra ismeretlen nyelven van, míg RL_D a dokumentum olvashatóságát jelöli. A fenti formula specifikus paraméterek között van, melyek egy bizonyos nyelvcsoporthoz vannak hangolva. Ezekben benne van a dokumentumban szereplő összes szó és mondat, valamint a hat betűnél több karakterből álló szavak.

Bár ez a formula sikeres lehet az angol nyelv vagy a francia nyelv esetén, sajnos a többi nyelvénél már kevésbé hatékony, mivel a Nyugat-Európai nyelvekből nyert adatok túl szűkre szabják a paramétereket, mert azok a Nyugat-Európai nyelvekre lettek tervezve.

Miután a Laesbarhed-index formula függ attól, hogy egy szövegben hány hatnál több karakterből álló szó van, ezért eredménytelen olyan nyelvek esetében, melyek más írásrendszert alkalmaznak – ilyen pl. a kínai nyelv. Ez az alkalmazhatatlanság a kínaihoz, ill. a hozzá hasonló nyelvekben azért van, mert itt az írásjelek fogalmakat jelölnek betűk helyett, valamint a kínaiiban a szavak gyakran csak egy vagy két betűből állnak a szemantikai összetettségétől függetlenül. Hasonló gondolatot követve, sok nyelv (még azok is, melyek írásban betű alapokon nyugszanak) nem tartja be a Laesbarhed-index azon feltételezését, hogy a szemantikailag komplexebb szavak több karakterből állnak, mint a kevésbé összetettek. Tehát ezekben a nyelvekben nem lehet ugyan azt a mércét használni a dokumentumok olvashatósági szintjének meghatározására, mert a szavak hossza nem egyezik meg azok jelentéstartalmi komplexitásával (Kirchner, Nuger és Zhang, 2009.).

Már az elején meg kell jegyezni, hogy a gépi fordítás nem feltétlenül a legjobb módszer a többnyelvű információkeresés problémájának a megoldására, mivel elég messze van a lehető legjobb pontosság elérésétől, valamint a lehetséges és interneten elérhetővé vált hatalmas mennyiségű dokumentum-gyűjteményt teljesen lefordítani igencsak nagy kiadásokkal járna, emellett teljesen felesleges pluszfeladatok végrehajtásával is járna – például a szórend helyes kialakítása.

Ezért a mai kutatások inkább arra fókuszálnak, hogy magát a lekérdezést fordítsák le a célnyelvre. Ez sokkal költséghatékonyabb, tekintve, hogy a dokumentumok relatíve kis százaléka felé mutatnak nagy érdeklődést a felhasználók, ezért elég csak ezeket a dokumentumokat lefordítani. Ezen felül a felhasználók egy részének van akkora szókinccse az adott nyelvben, hogy megértse a dokumentumot, bár magát a lekérdezést nem tudta volna helyesen megfogalmazni.

2.1 Szabály alapú gépi fordítás

A szabály alapú gépi fordításnál három fő mód van a probléma megközelítésére: közvetlen, közvetítő nyelv és áthelyezés.

- A közvetlen módszernél – melyet a korai gépi fordítók is használtak – a forrásnyelvi szavakat lecserélik a célnyelvi szavakra.
- A közvetítő nyelvnél nyelvektől függetlenül próbálják bemutatni a forrásnyelvi szöveg elemzését, majd ebből képzik a célnyelvet.
- Az áthelyezésnél a forrásnyelvi szöveget analizálják, majd meghatározott szabályok szerint feltérképezik a forrásnyelvi szöveget és azt átültetik a célnyelvibe (*Prósztéky*, 2005.).

A szabály alapú gépi fordítás általában a Vauquois-háromszöget alkalmazza (*Vauquois*, 1968.). Ebben négy különböző szintet állapíthatunk meg: a legalacsonyabb szinten a forrásnyelv mondatait szóról szóra lefordítják a célnyelvre. Itt alapvetően az történik, hogy egy kétnyelvű szótár alapján lefordítják a szavakat azok lehető legjobb megfelelőire a másik nyelven. Némi módosítás lehetséges a szavakon vagy azok sorrendjén, hogy jobban illeszkedjen a célnyelv nyelvtanához (*Nie*, 2010.).

Világosan belátható, hogy ez a könnyed megközelítés mennyire képtelen megbirkózni a természetes nyelvek összetettségével. Hiszen nem mindig lehet egy kifejezést lefordítani, valamint a szavak jelentése erősen függ a kontextustól is. A második és harmadik szint átvitelt használ: előbb szintaktikait, majd szemantikait. Ez a megközelítés előbb elemzi a forrásnyelv mondatait, hogy felismerje a szavak szintaktikai és szemantikai szerkezetét, majd ezt behelyezi a célnyelvre. Végezetül létrejön egy célnyelvi mondat, ahol a lefordított szavakat és kifejezéseket helyes szerkezetbe teszik. A negyedik szint a közvetítő nyelvi megközelítés, ahol nyelvtől függetlenül próbálják megjeleníteni a mondatot. Ezután a célnyelvi fordításnak már csak erre a közvetítő nyelvre kell utalnia.

Elméletben a közvetítő nyelvi megoldásnak több előnye van az áthelyezővel szemben. Nem kell különleges áthelyező szabályokat létrehozni minden egyes nyelvpárra, hanem elég fordítania a nyelvek és a közvetítő nyelv között. Bár azt gondolná az ember, hogy egy ilyen megoldás képes lenne megjeleníteni mondatokat minden nyelvben, azonban kiderült, hogy egy ilyen rendszer létrehozása igencsak komoly nehézségekbe ütközik (Nie, 2010.).

2.2 Statisztikán alapuló gépi fordítás

Ez a fordításfajta a párhuzamos gyűjtemények fordítása során keletkezett példákon alapul. Egy ilyen gyűjteményt tovább lehet bontani hozzárendelt mondatokra. Az ilyen párhuzamos szövegekben a különböző típusú kapcsolatokat ki lehet nyerni és azokat felhasználni új szövegek lefordításához. A kifinomultabb rendszerekben a szavak közti fordítási kapcsolatokat nem fordítják le a forrásnyelvi és célnyelvi mondatban elfoglalt helyzetétől függetlenül. A szavak sorrendje is felcserélődhet – akár kifejezésekben is. Például magyarban „naprendszer”-t írunk, míg a franciában „système solaire”-t (Nie, 2010.). Ezt a faktort torzulásnak nevezzük. Emellett egy szót egy másik nyelvben akár több másikra is le lehet fordítani (pl. „pomme de terre”, ami franciául a burgonyát jelenti). Ezt pedig termékenységnek nevezzük.

A mai tanulmányokban már a kifejezéseket is vizsgálják és arra az eredményre jutottak a kutatók, hogy a kifejezések szóról szóra való lefordítása lehetetlen gépi fordítással, mert egységeként kell őket kezelni, és ha egy kifejezést nem tudunk egyben lefordítani, akkor az azt alkotó részek lefordításával nem tudjuk újból létrehozni.

A statisztikai fordításban szükségünk van egy dekódolóra, hogy megállapíthassuk a lehető legjobb szórendet a célnyelvben. A sugárkeresés és a Viterbi keresés két ilyen megközelítés. Többnyelvű információkeresés során nem célunk a helyes szórend elérése a célnyelvben, hanem a legmegfelelőbb fordítási kifejezések megtalálása az.

A statisztikai fordítás során a legjobb találat elve érvényesül, mert bár a legideálisabb esetben azt szeretnénk, hogy minden kifejezést párosítani tudjunk a dokumentumban, ennek az esélye sajnos kicsi (Sparck Jones, 1998.). Azoknál a dokumentumoknál, amelyeknél minden kifejezést párosítani tudunk, nagy valószínűség szerint minden találat releváns lesz a keresés szempontjából, de viszont lesznek olyan dokumentumok is lesznek, amelyeknél egyáltalán nincs egyezés, ám mégis relevánsak lennének. A teljes egyezés hiánya magából a helyzetből adódik: a kifejezést részlegesen fejtik ki, mert a felhasználó nem tud többet vagy

jobban megfogalmazni. Esetleg nem közvetlen módon férnek hozzá a tartalomhoz, mert a szavak csak egy bizonyos mértékig írják le azt. Lehetséges az a helyzet is, hogy tartalom különböző olvasata a felhasználók között is hozzájárul a teljes egyezés hiányához (*Sparck Jones, 1998.*).

3. Tudás alapú technikák

Ezek a technikák többnyelvű szótárakat, -teauruszokat és -ontológiákat alkalmaznak a többnyelvű információkeresés feladatának végrehajtására. Már a legelső próbálkozások is szótárak felhasználásával kísérelték meg megoldani a többnyelvű információkeresést. Bizonyított tény, hogy a szótár alapú lekérdezés-fordítás, amelyekben minden kifejezést és fogalmat a másik nyelvben lévő megfelelőjével összepárosítanak – vagy ha ez nem lehetséges, akkor annak egy értelmében legközelebb álló, releváns szóval kapcsolják össze –, megfelelő első kísérlet a többnyelvű információkeresésre. Természetes azonban, hogy ezek a megoldások mindig gyengébben teljesítenek, mint az egynyelvű keresések. A kutatások szerint a gép által olvasható automata lekérdezés-fordítás az egynyelvű keresés pontosságának csupán negyven-hatvan százalékát éri el (*Hull és Grefenstette, 1996.; Ballesteros és Croft, 1996.*). Ennek három oka van:

- az első, hogy az átlagos szótárakban nincsenek meg a szakszótárak szavai, kifejezései,
- a második a hamis, illetve rossz fordítások jelenléte,
- a harmadik, hogy nem tudják lefordítani a többszavas kifejezéseket.

Az EMIR-rel (European Multilingual Information Retrieval, azaz Európai Többnyelvű Információkeresés) azonban sokkal jobb eredményeket értek el (*Peters, 1997.*). Az EMIR bebizonyította a többnyelvű adatbázisokban tett többnyelvű lekérdezések megvalósíthatóságát, beleértve a többnyelvű dokumentumokat is. A kísérlet egy rangsorolós Boole-i műveleteket (*Ungváry és Vajda, 2002.*) használ kétnyelvű kifejezés- és szólás-mondás szótárakkal együtt. Fontos megjegyezni, hogy egy tárgykörhöz tartozó fogalomszótár és kiterjedt manuális szerkesztés szükséges ahhoz, hogy el lehessen érni ezt a teljesítményt.

Ballesteros és Croft (1997) munkájukban bemutatják, hogy hogyan lehetséges fordítás előtti és utáni helyi kontextus vizsgálatokat használva a lekérdezések bővítése során jelentősen csökkenteni a hibákat a szótári fordításban és, hogy hogyan lehet segíteni a többszavas kifejezések pontos lefordítását. További vizsgálatokat is folytatnak, hogy egy

gyűjtemény alapú többnyelvű összepárosító teaurusz megoldja-e azt a problémát, hogy a szótári alapú fordításban nincs elég kontextus a kifejezések pontos fordítására.

3.1 Teaurusz használata

Számos kísérlet és a gyakorlat is bebizonyította, hogy a teaurusz alapú rendszerek rendelkeznek az egyik legjobb pontossági aránnyal a többnyelvű információkeresés területén. Maga a teaurusz egy olyan ontológia, amely terminológiákra specializálódik, és fogalmi kapcsolatokat biztosít. A többnyelvű teauruszt értelemszerűen több nyelv esetén alkalmazunk. A teaurusz, a benne lévő osztályozás segítségével sok funkciót lát el, melyek mindegyikét nagyban befolyásolja, ha mindezt a többnyelvű környezetben használják (Soergel, 1997.). Az ISO 25964-1-es (korábban az 1985-ös, 5964-es) szabványa írja elő, hogy hogyan építsük be egy tárgykör tartalmát egy többnyelvű teauruszba és meghatározza az alternatív technikákat is. A teauruszok hátránya, hogy kialakításuk és a fenntartásuk költséges, valamint képzés szükséges a megfelelő alkalmazáshoz is.

Soergel (1997) szerint két módja van a többnyelvű teauruszok felhasználásának:

- az egyik, hogy ellenőrzött szótári hivatkozást és keresést lehet bennük végezni,
- a másik, hogy tudás alapú támogatást ad a szabad szöveges keresés alá.

Az ellenőrzött szótári hivatkozásos és kereséses többnyelvű teauruszt olyan egynyelvű teauruszok csoportjának lehet elképzelni, melyek mind egy közös fogalmi rendszert térképeznek fel. Egy ilyen szótárban meghatározott számú kifejezés van, amire lehet hivatkozni és keresni. A nyelveken átívelő keresés célja, hogy felhasználó olyan dokumentumok között is kereshessen, melyek nyelvét nem beszéli, s amelyekre a saját nyelvén feltett kereső-kifejezésekkel próbál meg rátalálni. Ennek legegyszerűbb módja, hogy a felhasználó egy teauruszban megtekinti egy kifejezés célnyelvi megfelelőjét.

Az ellenőrzött szótárak problémája, hogy a szótárban lévő szavakat muszáj hozzárendelni a gyűjteményben a megfelelő dokumentumhoz. Ezt tradicionálisan manuálisan viszik végbe, de már törekednek arra, hogy a folyamatot gépesítsék. Egy másik gond, hogy a felhasználókat meglehetősen nehéz megtanítani a hatékony használatra (Ungváry és Vajda, 2002.). Előnye viszont, hogy közös forrásból jönnek a szavak és kifejezések, ami csökkenti a bizonytalanságot, valamint hála a manuális kiválasztásnak, teljesen médiafüggő (Vet és Bakel, 1998.).

A szabad szöveges keresés egy sokkal összetettebb feladat. Ez megköveteli, hogy a lekérdezésben szereplő összes lekérdező-kifejezés fel legyen térképezve a szöveg nyelvén lévő kereső-kifejezésekre, melyek lehetőleg olyan információkkal vannak ellátva, hogy az adott kereső-kifejezés elárulja, hogy mekkora mértékben járul hozzá a szövegben a kereső-kifejezés előfordulása a lekérdező-kifejezés relevanciájához (Peters, 1997.). Soergel (1997) szerint a szabad szöveges keresés legnagyobb problémája abból fakad, hogy az tényleges használat közben dolgozik, míg az ellenőrzött szótárnál lehetősége van a felhasználónak, hogy befolyásolja a használatot. A szabad szöveges keresés lekérdezési potenciálja viszont nagyobb, mint a szótáras rendszeré.

Az ontológiák és tezauruszok legfőbb problémája, hogy drága a kiépítésük és a fenntartásuk, ráadásul nehéz őket frissíteni. A nyelvek közti különbségek és a kulturális tényezők miatt nehéz a hatékony feltérképezés elérése a lexika és fogalmi megfelelők között a két adott nyelvben, ezt a problémát tovább súlyosbítja, ha több nyelvvel kell dolgozni. Ezért elengedhetetlen valamilyen közvetítő nyelv, mely lehetővé teszi a többi nyelv közötti fordítást, ennek velejárója az lesz, hogy kompromisszumot kell kötni a többnyelvűség és az egynyelvű specializálódás között (Peters, 1997.).

4. Szótár alapú fordítás

Egy kétnyelvű szótárban minden szó vagy kifejezés egyenként le van fordítva a célnyelvre, gyakran több szinonimával együtt. A szótárak különböző elvek alapján vannak felépítve. A többnyelvű információkeresésnél használatos szótárak általában szólisták, melyek mellett ott van a fordításuk is. Némely szótárban példák vannak a felhasználásra, vagy jelentésmagyarázattal is kiegészítik a szavakat. A szótárban lévő többletinformáció segíthet kiválasztani a legmegfelelőbb fordítást a szöveg kontextusából. Azonban a legtöbb tanulmányban az ilyen többletinformációk nem elérhetőek a felhasznált szótárakban (Nie, 2010.).

A szótár alapú fordításban általában egy gép által olvasott kétnyelvű szótárt használnak, hogy lefordítsák a célnyelv és a forrásnyelv megegyező jelentésű szavait, kifejezéseit. Viszont itt van néhány probléma jelentkezik: pl. vannak olyan szavak, amik nincsenek a szótárban. Nincs olyan szótár, amely teljes lenne, hisz a használatban lévő nyelvek fejlődnek, módosulnak – egyes szavak több jelentést is kaphatnak, de akár teljesen

megváltozhat a jelentésük. Emellett a több jelentésű szavak már önmagukban is gondokat okozhatnak, de megeshet még az is, hogy az adott szó több jelentéssel bír a célnyelvben is – ez növeli a fordítás bizonytalanságát és ezzel párhuzamosan csökkenti a pontos keresés esélyét (*Nie, Simard, Isabelle és Durand, 1999.*).

Probléma lehet az is, hogy esetleg nincs megfelelő szó vagy kifejezés az adott célnyelvben. További nehézségek: a szótárakban általában a szavak alapesete áll fent, míg a keresésben lehet, hogy azok egy változatát írják be; a kifejezéseket előbb fel kell ismerni, hisz lehet, hogy csak hasonló jelentéssel bír célnyelv kifejezésével, ezért fel kell őket bontani, hogy kereshetővé lehessenek (*Hedlund, Järvelin, Keskustalo és Heikki, 2001.*). Ez utóbbi két gond fennáll a hagyományos keresőrendszereknél is, így itt is alkalmazható az ott használt megoldások egyike.

Bár a szótár alapú gépi fordításnál rengeteg rendszer tűnik ígéretesnek, azonban nagy részüknél nem lehet eléggé kibővíteni a szótár méretét úgy, hogy az ne okozzon problémát a felhasználóknak. Erre egy lehetséges megoldás, ha maga a felhasználó tudja irányítani az új szavak vagy mondatnyi szerkezetek hozzáadását, illetve elvételét (*Prószéky, 2005.*). Ennek segítségével egy képzett felhasználó úgy képes hozzáadni és megváltoztatni nyelvtani szabályokat, hogy közben nem kell lefordítani a többi szót.

4.1 Alap megközelítés

A szótárakat általában szóról szóra fordításra használják. Amikor egy forrásnyelvi lekérdezést szeretnénk lefordítani, fel kell tennünk magunknak a kérdést, hogy melyik fordítás a leghelyénvalóbb és melyiket válasszuk ki. Mint már említettük, a legtöbb szótár nem tartalmaz többletinformációt, ami segítene e folyamatban. Ilyen esetekben a korai tanulmányok két módszert javasoltak: a lekérdezés szavára felhasználjuk az összes lehetséges fordítást, vagy az első fordítást használjuk, amely szerepel a szótárban (*Nie, 2010.*).

Az első módszernél minden lehetséges kifejezést fel lehet használni a célnyelvben, s így kibővíthetővé válik a lekérdezés. Cserébe viszont fokozódik a bizonytalanság és megnöveljük a fordítás pontatlanságát. Emellett rengeteg szónak több jelentése is van, és ha ezek szerepelnek egy szótárban, akkor általában összhangban is vannak a jelentéseikkel. Ezért ha az összes lehetséges fordítást felhasználjuk, elkerülhetetlenül rossz eredményeket is kapunk, mivel a kontextust is figyelembe kell venni. Ez pedig irreleváns dokumentumok megtalálásához vezet. Kijelenthetjük, hogy ez a fajta stratégia nem hatékony.

A második módszer alapelve, hogy általában egy szó első fordítása a leggyakrabban használt fordítás is egyben – persze ez függ a szótár rendezőelvétől is. Így elviekben nagyobb az esély a helyes fordítás eltalálására. Ez a technika hasonlatos a gépi keresés alapértelmezett működéséhez, amikor is nincs további elérhető információ a szóról. Ennek a módszernek a legnagyobb hátránya, hogy szótárfüggő, hogy melyik szót használják az első fordításra, és így szótáronként más-más eredményt kaphatunk.

Mindkét módszer elég gyengén teljesít az egynyelvű kereséshez képest – alig felét éri el a hatékonyságának. Ez alapján kijelenthető, hogy az ilyen egyszerű módszerek alkalmatlanok a többnyelvű keresésre (Nie, 2010.).

4.2 Kifejezések súlyozása

Mint láthattuk az egyszerű módszerekkel több probléma is van, ezért is fontos a szavak és kifejezések súlyozása. Ez főleg annál a módszernél fontos, amikor minden lehetséges fordítást felhasználunk, hisz ezekből sok nem is releváns, így ki kell választanunk, hogy melyiket is érdemes alkalmazni a továbbiakban. Emellett előfordulhat az is, hogy egy kifejezésben egy másik, kevésbé releváns szóra kapunk több találatot a keresés során, ezért valahogy ezt ellensúlyozni kell. Erre egyszerű megoldás a kifejezések normalizálása, ilyenkor a kifejezések súly $1/n$ lesz, ahol az n a forráskifejezés fordításainak a száma. Xu és Weischedel bebizonyította (Xu és Weischedel, 2005.), hogy ez az egyszerű eljárás hatékonyan képes kijavítani a találati arányt. Sikerült a többnyelvű keresés hatékonyságát ötven százalékról, az egynyelvű keresés hetven-nyolcvan százalékára javítaniuk.

Pirkola és társai (1998) más szemlélettel álltak elő: egy szó különböző fordításait, annak szinonimáinak tekintették. #syn() operátort használtak az INQUERY rendszerben (Broglia, Callan és Croft, 1994.), amely szinonimaként tekint kifejezések egy csoportjára és összegyűjti a találatokat a csoportban. Egy ilyen szerkezetben a fordítási kifejezések kombinációja értelemszerűen megváltoztatja a lekérdezés szavainak súlyozását és lehetővé teszi a jobb egyensúly elérését a lekérdezés részei között. Kutatásaik során megmutatták, hogy a módszerük eléri az egynyelvű keresés hatékonyságának hetvenhét százalékát, míg a sima fordítás csak alig ötvenkét százalékos hatékonyságot tudott felmutatni.

4.3 A szótárak terjedelme

A fordítás minősége erősen függ a szótár minőségétől, beleértve a helyességet és a teljességet (vagy lefedettséget). Azonban nehéz letesztelni, hogy mekkora e hatás nagysága, mivel már az is nehéz feladat, hogy eldöntsük, milyen szempont szerint vizsgáljuk a szótárat, emellett megmérni szintén bonyolult dolog. Ennek ellenére Xu és Weischedel (2005) megvizsgáltak egy aspektust: a forráskifejezések lefedettségét a szótárban. Megpróbálták különböző méretűre csökkenteni egy szótárat (a forrásnyelvi kifejezések tekintetében) a leggyakoribb elemek megtartásával és megvizsgálták ennek hatását a keresés hatékonyságára vonatkozóan. Azt az eredményt kapták, hogy egy bizonyos mértékig nő a keresés hatékonysága, onnantól viszont nyelvfüggő, hogy tovább növekszik-e. Bár nem nézték, hogy mennyire teljes a forrásnyelvi kifejezések fordítása, azt viszont jogosan feltételezhetjük, hogy egy teljesebb szótár általánosságban javíthatja a keresés hatékonyságát. Ellenben, amikor a fordítások száma megnövekszik, akkor ezzel párhuzamosan a bizonytalanság is emeli.

Mindkét aspektust figyelembe véve, azt várhatjuk, hogy javul a többnyelvű keresés hatékonysága mind a forrásnyelv mind a cél nyelv szavai között. Egy másik dolog, amit még nem vizsgáltak, az maga a szótárak fordításainak a helyessége, pedig a keresés hatékonysága ezen áll vagy bukik.

4.4 Fordítási bizonytalanság

Azokban a kísérletekben, ahol egyszerű megoldásokkal próbáltak operálni, a szótárakkal alapvető problémák voltak (Ballesteros és Croft, 1997.):

- Bizonyos kifejezések vagy nem voltak meg a szótárakban, vagy a fordításuk nem volt naprakész.
- A fordítás helyessége eleve bizonytalan volt.
- Némely kifejezést nem lehetett helyesen lefordítani, mert a szótár nem fedte le az összes lehetséges előfordulást.

E problémák közül az azonosított bizonytalanság és a hiányzó fordítások a legjelentősebbek szótárak használatánál. A hiányzó fordításokat meg lehet oldani fordítási kapcsolatok automatikus feltárásával. A következő részben a bizonytalanság feloldásáról lesz szó.

4.5 A fordított szavak kiválasztása

Tételezzük fel, hogy a helyes fordítás benne van a szótárban, több más lehetséges jelölttel együtt (akár azonos, akár más jelentéssel). A probléma abból áll, hogy ki kell választani a helyes (avagy a helyzetnek legmegfelelőbbet) fordítást vagy fordításokat a jelöltek közül. Az első, avagy a leggyakrabban használt fordítás kiválasztása az első lépés. De nem biztos, hogy ez majd megfelel a mondat kontextusának. Egy jobb megoldás, ha kontextus függővé tesszük a kiválasztást, például a lekérdezésben lévő szavak együttes előfordulása.

Grefenstette (1999) megfigyelte, hogy a helyes fordításoknak nagyobb a gyakorisága más együttes előfordulású fordításokkal. Ezért javasolta, hogy ezt használják a probléma megoldásához. Mivel valószínűtlen, hogy a helytelen fordításoknak nagy az együttes előfordulása, ezért feltételezhető, hogy a helyes fordítás az lesz, amikor az együttes előfordulás a legnagyobb.

Meg lehet próbálni helyes szórend létrehozását is a fordítások kombinálása során, mint ahogyan az Jang és társai tették (*Jang, Myaeng és Park*, 1999.). Ahelyett, hogy az összes lehetséges kombinációt figyelembe vették volna a fordításban, lekorlátozták a lekérdezésben lévő azt az egymást követő szavakra. Viszont ez sem egy tökéletesen biztos módszer, mert hasznos kombinációs párok maradhatnak ki a fordítási folyamatból.

Gao és társai által javasolt metódus az, hogy statisztikai elemzést végeznek, hogy megkapják a legerősebb kapcsolatot a lekérdezésben lévő szavak között, s ez nem feltétlenül követné a szórendet. Viszont az még nem világos, hogy a rövid lekérdezéseknél a kapcsolatok függését kiválasztani hatékonyabb-e, mint az összes lehetséges fordítást figyelembe venni. *Metzler és Croft* (2005) sikeresen egyesítettek független kapcsolatokat a lekérdezésbe: mind a rendezett, mind a rendezetlen függőségeket egyesítették a szomszédos szavakkal egy Markov-mezőben, és arra jutottak, hogy ez a modell jelentősen felülmúlja azokat a modelleket, amelyek nem veszik figyelembe a kifejezések függőségét.

A lekérdezések szintaktikus szerkezetét is fel lehet használni a fordítási folyamatban: lehet, hogy jobb megvizsgálni azoknak a szavaknak a kohézióját, amelyek főnévi kifejezést alkotnak a lekérdezésben. A szintaktikus megkötés is mutatott némi javulást a többnyelvű keresés hatékonyságában, azonban ez még mindig kisebb, amikor a statisztikai kapcsolatokat is számba veszik. Ennek valószínűleg az az oka, hogy egyrészt a szintaktikai kapcsolat, amelyet észreveszünk, nem mindig pontos, másrészt, hogy a kontextusban lévő szavaknak

nem kell valamilyen különleges szintaktikai kapcsolatban lenniük ahhoz, hogy hasznosak lehessenek – ezt több tanulmányban is bebizonyították már.

5. Párhuzamos és összehasonlító gyűjteményeken alapuló fordítás

Ebben a típusban a fordítást egy már meglévő párhuzamos vagy összehasonlító gyűjteményből nyerjük, ennek a módszernek köszönhetően a pontosság is nagy. Azonban a párhuzamos gyűjtemények elég ritkák, így leginkább összehasonlítóakat kell használni. Az összehasonlító gyűjteményben a szavak nem egymás megfelelői, inkább csak témájukban hasonlítanak – pl. világgazdasági újságok, melyeket más nyelven adtak ki, de a hírek egy része ugyanaz az egyikben, mint a másikban. A forrásdokumentum és céldokumentum ebben az esetben felcserélhetők. Mivel nincs rá esély, hogy minden témának legyen egy megfelelő párja, ezért a gyűjtemény kisebb lesz, mint a forrásdokumentumok száma. Viszont megjegyezhetjük, hogy nem feltétlenül muszáj ezeknek a dokumentumoknak párban lenniük, mivel a forrásdokumentumot további hasonló céldokumentumokkal párosíthatjuk (Nie, 2010.).

A párhuzamos szövegeknél a szöveg fordítása egy másik nyelvvel együtt van jelen (Talvensaari, 2008.). Napjainkra egyre több és több párhuzamos szöveg érhető el. E szövegek gazdag forrásai a fordítási kapcsolatoknak – minden szinten: szövegek, mondatok, kifejezések és szavak szintjén. Sokféle megoldás született ezek kinyerésére. Érdekes módon a párhuzamos gyűjteményeknél alkalmazott módszerek jó része hasonlít a statisztikai gépi fordítás technikáira – tekintve, hogy a fordítási kapcsolatok gyakoriságát vizsgálják a gyűjteményekben (Nie, 2003.). Az összehasonlító szövegek témakörben hasonlatosak a szövegek, amelyek mégsem párhuzamosak. Összehasonlító szövegből több érhető el, mint párhuzamosból. Annak ellenére, hogy nehéz használni az összehasonlító szövegeket gépi fordításra, mégis használható megközelítés a többnyelvű információkeresésre, köszönhetően kevésbé szigorú követelményeknek.

A gyűjtemény alapú keresés megközelítés során nagy mennyiségű szöveget kell átvizsgálni, és az abból automatikusan kinyert információt kell felhasználni az alkalmazás-specifikus fordítási technikák felépítéséhez. A gyűjtemények állhatnak párhuzamos (párosító-megfeleltető) vagy összehasonlító (tárgykör-specifikus) dokumentumokból.

5.1 Párhuzamos gyűjtemények

Párhuzamos szövegeket már az 1990-es évektől alkalmaznak többnyelvű információkeresésre. Példa erre a kanadai Hansard vagy az Európai Unió EuroParl – ezek a parlamenti vitákat tartalmazzák, különböző nyelveken (Nie, 2010.). A nyers párhuzamos szövegek a szöveg szintjén vannak összepárosítva – a szöveg a fordításához van kötve. Ez a szint túlságosan is „durva” ahhoz, hogy közvetlenül felhasználhassuk a fordítási folyamatban. A legtöbb módszer először magasabb szinten próbálkozik meg a párhuzam felállításával, bekezdés és mondat szintjén, majd csak azután jön a kifejezések és szavak szintje. Erre azért van szükség, hogy csökkentsük a területet (azaz a szövegméretet), amelyen a szavaknak egyezniük kell. Ez lehetővé teszi, hogy jobban ki lehessen használni a párhuzamos gyűjteményeket és pontosabb fordítási modellt eredményez (Nie, 2003.).

Egy másik fontosabb modell az IBM modellek családja, melyre sok mai statisztikai rendszer épül. A rendszer Brown és társai munkájának az eredménye (1993). Ennek a módszernek az alapja, hogy megpróbálja megállapítani, hogy mely célnyelvi mondat felel meg a legjobban a forrásnyelvi mondatnak. Ebben a rendszerben nem veszik figyelembe a szavak sorrendjét, hanem azok csak egy halmazt alkotnak. Bármely szó, amely egyezést mutat a célnyelvi mondattal, lehetséges fordításként kezelik bármely forrásnyelvi szóra.

5.2 Bekezdés és mondat egyeztetése

A jó minőségű fordítások általában ugyan azt a bekezdés és mondat rendet követik. Az a bekezdés vagy mondat, amely először jelenik meg a forrásnyelvben, általában az lesz az első a fordításban. Vessünk egy pillantást Gale és Church (1993) algoritmusára, mely az egyik első ilyen technika volt.

Jellemző, hogy különleges angol-német-francia nyelvű formázó jelzőket használtak munkájukban (a svájci UBS bank dokumentumaiból dolgoztak). Ez az első lépés viszonylag könnyen és megbízhatóan ment, tekintve, hogy a gyűjteményben tökéletesen tisztán kivehetőek voltak a bekezdések határai és feltételezték, hogy egy az egyhez vannak összekapcsolva. A második lépésben a bekezdésben lévő mondatokat kapcsolták össze, itt már nem alkalmazhatták az egy az egyhez elvet, mivel egy mondatot akár több mondatként is le lehet fordítani (1-n kapcsolás), valamint fordítva is igaz a dolog, azaz több mondatot akár egy mondatra is le lehet fordítani (n-1 kapcsolás). Emellett egy mondatot ki lehet hagyni a

folyamatban (1-0 kapcsolás), valamint hozzá is lehet adni (0-1 kapcsolás). Hogy megbirkózzanak ezekkel a problémákkal, beleértve a mondatok általános sorrendjét a forrás- és a célnyelv között, *Gale és Church* (1993) azt az általános megfigyelést használta föl, hogy a hosszú mondatok általában hosszú fordítással bírnak, míg a rövid mondatok röviddel. Ezért egy dinamikus programozást javasoltak, hogy megállapíthassák a lehető legjobb összekapcsolásokat a mondatok között.

A mondat-összekapcsolási problémát a következőképp lehetne megfogalmazni: adva van két szöveg, két különböző nyelven, melyekről feltételezzük, hogy felbonthatóak mondatokra. Az egymást követő mondatokat részekbe lehet tömöríteni, de lehetnek üres részek is. Sikeres összekapcsolás akkor lesz, ha a részek megegyeznek egymással és nincs kereszteződés a részek között (*Nie*, 2010.). Feltételezzük továbbá, hogy két rész közötti kapcsolás független attól a kontextustól, amelyben szerepelnek.

Két megközelítés létezik a részek meghatározására:

- egy a mondatok hosszán alapul,
- a másik a lexikai szabályokon (*Nie*, 2010.).

A mondatok hosszát lehet mérni karakterekben vagy szavakban (*Brown, Lai és Mercer*, 1991.). *Gale és Church* (1993) úgy találta, hogy a karakter alapú megoldás jobban teljesít a szóbelinél az UBS gyűjteményében. A közeli nyelvekben, mint az angol és a francia, a rokon szavakat is fel lehet használni a mondatok összekapcsolására. A rokon szavaknak ugyan az vagy hasonló a gyökerük a különböző nyelvekben – elég csak az *információ* szóra gondolni (*Talvensaari*, 2008.): spanyolul *información*, olaszul *informazioni*, angolul, franciául és németül pedig *information*. A közös tö, az *inform* mindegyikben megvan. *Simard és társai* (1992) szerint, ha két mondat között rokon szavakat találunk, akkor jó esély van rá, hogy kapcsolatban állnak (*Simard, Foster és Isabelle*, 1992.). Tehát a hosszúságon alapuló mérésnél a rokon szavak biztosan megnövelik az összekapcsolásokat, de kritériumok hozzáadásával még tovább lehet növelni a kapcsolódást.

Ezt úgy lehet elérni, hogy egyéb nyelvi erőforrásokat vonunk be a folyamatba. *Kay* és munkatársai szótárak használatát ajánlották, hogy lexikai eszközök is rendelkezésre állhassanak (*Kay és Röscheisen*, 1993.). Ez a nyelvileg távolabb álló nyelveknél hasznos (pl. angol-kínai).

5.3 Összehasonlító gyűjtemények

Nem minden nyelvpárnál lehet alkalmazni, viszont sokkal könnyebb ezt használni, amikor a két nyelv szövegei ugyan arról a témáról szólnak, anélkül, hogy szorosan párhuzamban állnának. Erre remek példa a hírek témaköre. Egy-egy nagyobb hír szinte mindenhol megjelenik a világban, és bár nem párhuzamos bennük a szöveg, mégis elvárhatjuk, hogy bizonyos fontos elemek megismétlődnek bennük, ezért a tartalmuk összehasonlítható. Például *Sheridan* és *Ballerini* (1996) e módszer használatával gyűjtött összehasonlítható gyűjteményeket.

Tekintve, hogy az összehasonlítandó szövegek nem egységesek, sokszor „zavarosabbak”-nak is nevezik őket, nem lehetséges a párhuzamos szövegeknél alkalmazott technikák hatékony használata, hanem egy sokkal nyersebb megoldást kell találni. Általánosságban elmondható, hogy a kutatások azt próbálták megállapítani, hogy a nyelvek közti hasonlóság hogyan függ a kifejezések együttes előfordulásától (*Nie*, 2010.). Minél több kifejezés fordul elő együtt egyszerre a két szövegben, annál inkább feltételezhetjük azt, hogy hasonló a jelentésük. Többfajta kritérium alapján lehet a jelentést összehasonlítani, pl. koszinusz hasonlóság, közös információtartalom, stb. Érdekesség, hogy egy ilyen többnyelvű teaurusz, hasonló az egynyelvű teauruszhoz, azzal a különbséggel, hogy az együttes előfordulások két különböző nyelvben találhatók meg. Amikor elkészült egy többnyelvű teaurusz, akkor a lekérdezés fordításának folyamata rendkívül közel áll az egynyelvű keresés kibővítésének folyamatához: minden lekérdezett kifejezés súlyozva lesz hasonló kifejezésekkel egy másik nyelven.

Első pillantásra kockázatosnak tűnhet összehasonlító gyűjteményt használni többnyelvű információkeresésre, azonban a hatékonyság összhangban van az álrelevanciás visszajelzéssel és az egynyelvű keresés együttes előfordulásának elemzésével. Mivel azt szeretnénk itt is elérni, hogy a forrásnyelvi kifejezéseink ugyan azt jelentsék vagy hasonló jelentéssel bírjanak, mint a célnyelvi kifejezések, ezért ez a technika is használható van a többnyelvű információkeresésben (*Nie*, 2010.).

Nem szabad elsiklanunk azonban a felett, hogy egy ily módon létrejött teaurusz eléggé következtelen lehet, tekintve, hogy vannak kifejezések, melyek együttesen előfordulnak, mégis más a jelentésük. *Franz* és társai (1999) olyan megközelítést használtak, hogy megpróbálták a durván összehasonlított szövegeket jobban összehangolni kisebb részekkel. Ezek a szövegegységek nagyjából ötven szóból álltak. Egy bizonyos időkereten

belül megpróbálták meghatározni a lehetséges párhuzamos részleteket, úgy hogy egy kezdetleges fordítási forrással (szótár) lefordították a részletet és megnézték mikor érik el a legjobb eredményt, s ezt a halmazt nevezték párhuzamos résznek. Ez a módszer színvonalban magasabb minőségű, egymást követő egységeket eredményezhet, és az ebből felépülő fordítási modell, jobb eredményeket érhet el.

Ezt az eljárást ki lehet egészíteni statisztikai fordítási modellekkel, hogy megállapíthassuk a párhuzamosság mértékét a szövegek között. Ezen felül olyan elemeket is használhatunk támpontnak, mint a tulajdonnevek. Ezek a kibővítések hasonlóak a rokon szavaknál alkalmazott technikákhoz.

Mindegy, hogy milyen módszert alkalmazunk, az összehasonlító gyűjtemények zavaros mivolta megmarad, ami továbbra is rizikóssá teszi, hogy statisztikai fordítási modellt építsünk ki belőle. Ehhez még az is hozzájön, hogy a hasonlósági kapcsolatok kevésbé pontosak, mint a párhuzamos gyűjtemények fordítási kapcsolatai. De ha nem áll rendelkezésünkre jobb forrás, akkor az összehasonlítás jó szolgálatot fog tenni nekünk (Nie, 2010.).

6. További módszerek, amikkel javítani lehet a többnyelvű információkeresést

6.1 Elő- és utófordításos kibővítés

Ahogy már rámutattunk az előzőekben, a lekérdezés fordítása és a teljes szöveg gépi fordítása közti különbség az, hogy a lekérdezést nem feltétlenül kell egyetlen szóval lefordítani, valamint nem kell szó szerint ugyan azt jelentenie elég, ha hasonló jelentésű szavakra fordítjuk le őket. Az elő- és utófordítás elve hasonló az álreleváns visszajelzés technikájához, azaz először a forrásnyelvben kutatnak dokumentumok után és az ezekből kinyert kifejezésekkel kutatnak a célnyelvben releváns dokumentumok iránt (Nie, 2010.). Több tanulmány is vizsgálta az elő- és utófordítás mikéntjét. *Ballesteros és Croft* (1997) arra jutottak, hogy mind a pontosságot, mind a visszahívást javítja ez a fajta keresés. *McNamee és Mayfield* (2002) pedig kimutatta, hogy az előfordítás hasznosabb, az utófordításnál, míg *Levow csoportja* ennek az ellenkezőjét (*Levow, Oard és Resnik*, 2005.). Ez az egymásnak

ellentmondó eredmény is mutatja, hogy erősen megoszlanak a kérdéssel kapcsolatos elméleti és empirikus vizsgálatok következtetései.

Annyi bizonyos, hogy mindkét módszer javítja a keresés hatékonyságát, de nem lehet kijelenteni, hogy az egyik módszer eredményesebb lenne a másikonál. A kibővítési folyamat hatása erősen függ a dokumentumgyűjteményektől. Míg az utófordításos kibővítés alapja, hogy abból a célnyelvi dokumentumcsoportból szerezzük be a dokumentumokat, amelyekben majd végül végrehajtjuk a keresést, addig az előfordításnál más dokumentumcsoportot kell felhasználni (Nie, 2010.).

A különbség a kereséses és a másik csoport között, hogy az utóbbiból olyan kifejezéseket is találhatunk, amelyek majd hasznunkra válnak a kereséses csoportban, hogy ha azonos a témájuk (Nie, 2010.). Ha viszont eltérő a témakör, akkor az előfordítás kártékony hatású, ezért meg kell bizonyosodnunk róla, hogy amikor ezt a technikát alkalmazzuk, azonosak a témák – ami plusz feladat.

6.2 Összefüggéstelen összekapcsolás

Rokon nyelvekben (például az egy nyelvcsaládban lévők) a kifejezések hasonlíthatnak egymásra és csak kis mértékben változnak nyelvenként. Ilyen például a skandináv nyelveknél (norvég, svéd, dán, stb.) fordul elő vagy az újlatin nyelveknél (spanyol, francia, olasz). Ezek között a nyelvek között a többnyelvű keresés anélkül is megvalósítható, hogy teljesen le kelljen fordítani a szavakat és elég csak az összefüggéstelen összekapcsolást használni (Nie, 2010.). Buckley és társai (2000) kihasználták az angol és a francia nyelv hasonlóságát és meglepő eredményre jutottak: ez a fajta keresés elérte az egynyelvű keresés hatékonyságának a felét, ami megfelel az egyszerű szótárhasználat sikerességi mutatójának. A technika flexibilitása azonban nem csak előny, hanem egyben hátrány is, mert nem biztos, hogy a hasonló szavak kapcsolatban állnak egymással. Ezért javasolt ezt a technikát csupán akkor használni, hogy ha nincs jobb elérhető megoldás (Nie, 2010.).

6.3 Egyesítő fordítások

A fordítási erőforrásokat gyakran kombinálják a jobb eredmény elérése céljából. Lu és társai (2004) például a statisztikai fordítási modellt egyesítették egy szótárral. A kísérleteik

(angol-kínai közti keresés) eredménye igazolta az elvárásokat: a statisztikai modell önmagában tizenöt százalékos hatékonyságot tudott felmutatni, a szótár pedig tizennégyet, de a kettő együtt huszonhárom és fél százalékot. Más kutatók további nyelveken is megismételték a kísérleteket és náluk is hasonló javulást mutattak ki.

A módszer lényege, hogy mindegyik technika során létrejön egy lista a fordításokból, majd ezeket összevetik és a közös szavakat használják föl. Emellett a fordítások irányát – forrásnyelv-célnyelv és célnyelv-forrásnyelv – is lehet egyesíteni. Ez segíthet a fordítás bizonytalanságát csökkenteni (*Ballesteros és Croft, 1998.*). Ez arra feltételezésre épül, hogy egy szónak lehetnek specifikus és általános fordításai is, főleg, hogy ha statisztikai modellt használunk. Ezekben az esetekben a specifikus fordítást kell alkalmazni, mert az általános növeli a fordítási bizonytalanságot.

Hogy megkülönböztessük az általános fordítást a specifikustól a következő feltevést alkalmazhatjuk: minél specifikusabb a lefordított szó, annál valószínűbb, hogy az eredeti a jelentést kapjuk a visszafordítás után, ill. minél általánosabb a kiindulás, annál valószínűbb, hogy több lehetséges fordítást kapunk (*Nie, 2010.*).

De nem csak szavak és kifejezések szintjén lehet használni ezt a módszert, hanem lekérdezéseknél és dokumentumoknál is. Itt is két felsorolás jön létre, melyeket összevetve egy végső listát kapunk.

A két irány összevetésével jobb eredményeket lehet kapni, mint ha csak egy irányba vizsgálnánk meg a fordításokat (*McCarley, 1999.*), viszont ki kell emelnünk, hogy ez egy költségesebb módszer.

6.4 Átívelő fordítás

Meglehet, hogy az angol és bármely más nyelvek között sok forrás áll rendelkezésünkre a fordításhoz, azonban közel sem ez a helyzet a többi nyelvpárnál (*Braschler, Peters és Schauble, 1999.*). Ezért természetesen adódik az a megoldás, amikor az angolt, mint közbeeső nyelvet használjuk, hogy kereshessünk két, nem sok forrással rendelkező nyelv között. Annak ellenére, hogy a direkt fordítás nem lehetséges, mégis tudunk ezzel a módszerrel releváns dokumentumokat találni. Nem meglepő módon, az átívelő fordítás kisebb hatékonyságú, mint a direkt fordítás, bár vannak kivételek (*Lehtokangas, Airio és Järvelin, 2004.*). Megközelítőleg negyvenöt százalékos eredményességet lehet kimutatni az egynyelvű kereséshez képest (*Nie, 2010.*).

Egy másik módszer a konvolúciós fordítási modell, amely statisztikai modelleket is felhasznál, csak harminchét százalékos a hatékonyságú (Nie, 2010.). Ez azért lehetséges, mert bár a lekérdezés fordításánál üdvös, hogy több lehetséges fordítás van, amivel kibővíthető a lekérdezés, de a dokumentum fordításánál már kevésbé az. Valójában a konvolúciós modellben a lefordított szavak válogatás anélkül mennek tovább, ezért a végső fordítás elég bizonytalan lesz.

„Háromszögelési” módszerrel is próbálkoztak növelni a hatékonyságot (Gollins és Sanderson, 2001.). Ennél az eljárásnál több nyelvet is felhasználtak közbeeső nyelvként az átívelő fordítás során. Kiderült, hogy így nem lehet javítani az eredményességet, bár később mások (Lehtokangas, Airio és Järvelin, 2004.) ehhez még használtak lekérdezéses fordítást is, amely segítségével már elérték a direkt fordítás szintjét.

Ez az eredmény különösen azért érdekes, mert megmutatja, milyen lehetőségek vannak ebben a módszerben, és hogy hogyan lehet növelni a keresés relevanciáját, amikor a dokumentum direkt fordítását nem lehet megoldani (Nie, 2010.).

6.5 Az egynyelvű és többnyelvű kapcsolatok egyesítése

Több tanulmány is kutatta (Gao, Nie, Xun, Zhang, Zhou és Huang, 2001.; Liu, Jin, és Chai, 2005.), hogy föl lehetne-e használni a többnyelvű információkeresésben az egy nyelvben előforduló kifejezések közötti kapcsolatokat. Arra jutottak, hogy igen, viszont az egynyelvű és többnyelvű kapcsolatokat külön kezelték. Lehetséges-e, hogy ha ezeket együtt kezeljük, akkor még jobb eredmény jönne ki?

Cao és társai arra jutottak, hogy nagyon is (Cao, Gao, Nie és Bai, 2007.). Ők maguk két egynyelvű kapcsolatot vettek figyelembe: a *tartalmaz* és az *együttes előfordulás*-t. Az előbbit arra használták, hogy összekössék a kifejezést, az azt alkotó szavakkal. Ez leginkább a kelet-ázsiai nyelveknél fontos, ahol a mondatokat fel lehet osztani rövidebb és hosszabb szavakra. A második kapcsolat pedig összekapcsolja a kifejezéseket, melyek gyakorta fordulnak elő együtt. A többnyelvű kapcsolatok a szótárakban lévő kapcsolatok voltak.

Az eredeti lekérdezések más-más úton kapcsolódtak a kifejezésekhez, és minél több kapcsolódott össze, akkor annál erősebb is volt a reláció. Cao-ék azt is kimutatták, hogy a helyes fordítások egyre több viszonyt hoznak létre, míg a helytelenek csak egyszeri kapcsolatot eredményeztek. Így ha a végére érünk a folyamatnak, akkor megkapjuk, mennyire is megfelelő a fordítás (Nie, 2010.).

A kapcsolatok egyesítésének technikája pontosabb eredményeket ad az elő- és utófordításos kibővítésnél – bár az is igaz, hogy az előbbi technika magába foglalja az utóbbi módszert. Kimondható hát, hogy érdemes lenne további kutatásokat végezni ebbe az irányba.

7. Többnyelvű információkeresés többnyelvű ontológiák felhasználásával, melyek a fordítás és a lekérdezés bővítéséül szolgálnak

Az internetfelhasználók körében az a növekedő igény, hogy többnyelvű környezetben keressenek információkat, kiemelt kutatási téma lett az információkeresés tudományában. Abusalah, Oakes és Tait (2009.), tanulmányukban többnyelvű ontológiák segítségével próbálják csökkenteni a fordítási bizonytalanságot. Értékelik a rendszerüket és összehasonlítják a gép által olvasható szótárak rendszereivel (Abusalah, Oakes és Tait, 2009.).

A többnyelvű információkeresés két ágra oszlik:

- az egyik a kétnyelvű, gépek által használt szótári rendszer és a gépi fordítás,
- a másik a koncepció (vagy tudás alapú) központú megközelítések (Abusalah, Oakes és Tait, 2009.).

A fő gond a kétnyelvű szótárakkal, hogy a fordítás pontossága bizonytalan, és emellé olyan problémák társulnak, hogy nehezen tud megbirkózni a ragozással, a vegyületek neveivel, a kifejezésekkel, a tulajdonnevekkel, a különböző helyesírási változatokhoz és a különleges szavakkal. A gépi fordítás általában ezt úgy próbálja megoldani, hogy a szavak szövegbeli kontextusát vizsgálja. Azonban egy átlagos keresőlekérdezésnél kevés a kulcsszó, így szinte nincs is mit vizsgálni. Ezért is hatékonyabbak teljes szövegek lefordításakor, mivel ott a szövegösszefüggés sokkal világosabb.

A koncepció központú megoldások, úgy, mint a tezaurusok és a többnyelvű ontológiák segíthetik áthidalni a nyelvi kifejezések és a jelentésük között lévő szakadékot. A kétnyelvű tezaurusz a hasonló jelentésű szavakat egy csoportba vonja össze, s osztály és tagozat szerint hierarchiát alakít ki benne, majd a jelentésük alapján térképezi fel őket (Nie, 2010.). Azonban a tezaurusok általában nem tartalmazzák egy szó definícióját, s valójában az egy csoportban lévő szavak csak „szimplán” kapcsolódnak egymáshoz. Emellett az egy vezéreszkriptor alatt lévő szavak, fogalmak különböző szintaktikai kategóriákba is eshetnek.

Az Abusalah, Oakes és Tait által készített tanulmány egy kétnyelvű ontológia létrehozását célozta meg, mivel úgy vélték, hogy az ontológia egy olyan általános

ismeretgyűjtemény, amelyet arra használnak majd, hogy kontextust adjon a lekérdezéseknek a lekérdezések bővítése során, így téve lehetővé a szavak fordításának a pontosítását. Az ontológia meghatározza a fogalmakat, kifejezéseket és a szókincset egy tárgykörben, valamint az e fogalmak között lévő kapcsolatot is. A fogalmak rendszertani szerkezetében az alosztályok megöröklik az előző osztály tulajdonságait és különállnak a „szuperosztályoktól”. A jelenlegi szemantikus web technológiák még arra is képesek, hogy új tényekre következtessenek az olyan régiekből, amelyek már részei egy ontológiának.

7.1 Az ontológia a gépi olvasású szótárral szemben

Abusalah, Oakes és Tait ontológiája arra épült, hogy a változó témaköröket modellezze, mely két ontológiból áll (arab és angol). Mindkét ontológia rendelkezik egy angol-arab kétnyelvű indexszel. Száz angol fogalmat párosítottak a nekik megfelelő arabbal, majd további száz kifejezéssel frissítették a mintát. Az automatikus ontológiai feltérképezés, mely szójelentés egyértelműsítést használt, nyolcvanhárom százalékos pontosságot ért el egy felhasználói felmérésben. Az olyan fogalmi kapcsolatok mellett, mint az „ez egy” vagy „van neki”, az ontológia olyat is használ, mint például „ez esetben”, stb.. A fogalmi kapcsolatokat arra használták, hogy szemantikailag összefüggő kapcsolatokkal kibővítsék a lekérdezéseket, és hogy javítsák az információkereső rendszer egynyelvű és nyelveken átívelő hatékonyságát (Abusalah, Oakes és Tait, 2009.).

Például a „hotel” az egy „szállás”, ezért ha a hotel egy kereső-kifejezés volt, akkor a lekérdezés kibővült a hotel és a szállás kifejezésekre, hogy több releváns találatot adjon az egynyelvű (angol) keresés során. Így az arab megfelelőikre is több releváns találatot adott a többnyelvű keresés folyamatában. Egy gépi szótárat is használtak arra az esetre, ha az ontológiai módszer nem működne. Mind az angol, mind az arab ontológia szemantikai szinten működött, s minden fogalomnak volt megfelelője a másik nyelvben.

A lekérdezés során teljes szövegkeresést alkalmaztak. Ez azt jelenti, hogy azokra a gyűjteményekre keresnek, amelyben a lekérdezés szavait összehasonlítják a benne szereplő dokumentum összes szavával. A gyűjteményben szereplő összes dokumentum számára egy hivatkozási indexet is létrehoztak a teljes szövegkereséshez. A gépi szótár alapja az Al-Mawred féle arab-angol szótár volt, mely kb. 100.000 angol-arab és 67.000 arab-angol bejegyzést tartalmaz.

Az ontológia alapú információkereső rendszer elküldi a kulcsszavakat az XSL-be (bővíthető stíluslap nyelv), hogy a lekérdezze az ontológiákat, kiválasztva a kapcsolódó fogalmakat és kapcsolatokat. A kifejezéseket szinonimákat keresve összehasonlítja, és ha talál akkor megkeresi a másik nyelvben azok megfelelőit a kétnyelvű index segítségével. Hogyha nem található meg egy fogalom az ontológiában, akkor a szótár a legrelevánsabb kifejezést keresi meg (*Abusalah, Oakes és Tait, 2009.*).

7.2 Értékelés

Abusalah, Oakes és Tait ontológiájának értékelése a használók fontossági sorrendjén alapul, melyeket kísérleti keresések alkalmával jegyeztek föl. Huszonöt közös angol nyelvű lekérdezést állapítottak meg a szakértőkkel történő beszélgetés során a turizmus és az utazás témakörében. A keresők, akik a mindkét rendszert értékelték, bár nem voltak jártasak a turizmusban, viszont legalább egyszer utaztak már az életükben. Mindegyikük anyanyelve arab volt, és jól beszélték az angolt is. Húszan voltak olyanok, akik minden egyes lekérdezést értékelték, ily módon ezer vélemény született.

Az egyszerű minősítés mindössze két pontból állt: releváns, ill. nem megfelelő találat. Ugyanakkor súlyozni is kellett a találatokat – nem releváns, nem tudja, inkább releváns, releváns, pontos (*Abusalah, Oakes és Tait, 2009.*).

A két kísérlet alkalmával, az elsőben a lekérdezés olyan szótári rendszerben folyt, melyek címmel és rövid leírással is rendelkeztek, emellett a módszer értékelte az első negyven találatot; a második esetben ugyan ez folyamat az ontológiai rendszer alkalmazásával folyt.

A felhasználók felmérése után az átlagos precizitás átlagát mérték, ahol az átlagos precizitás átlaga a releváns dokumentumok megtalálásának átlagos pontosságát jelenti. Az előbb említett kísérletekben a szótári rendszer negyvenkét százalékos találati pontosságot ért el, míg az ontológiai rendszer hatvanhárom százalékot, amely jelentősen meghaladja a szótári rendszer precizitását (*Abusalah, Oakes és Tait, 2009.*).

Bebizonyosodott tehát, hogy az ontológiai rendszer sokkal jobb eredményeket tud felmutatni, mint a szótári rendszer. Az ontológiai rendszer előnye, hogy nem csak a szóról szóra fordítás során lehet alkalmazni. Ezek az eredmények azért is különösen érdekesek, mert szemben állnak a korai egynyelvű kísérletekkel, melyekben a lekérdezés bővítése csökkentette a keresés hatékonyságát, mintsem növelte volna azt. Nehéz meghatározni, hogy

a fejlődés a lecsökkentett, illetve ismert témakör miatt jött létre avagy valamilyen másik ok áll a háttérben.

Mindkét rendszer értékelése során kiderült, hogy az ontológiai rendszer sokkal pontosabb, mint a tisztán szótári rendszer. A jövőbeli kísérletekben a szerzők megpróbálják növelni és kiterjeszteni az ontológiai rendszert, magyarázó jegyzetek segítségével, új fogalmakat rendelnek hozzá, hogy újra megmérthessék a szótári rendszerrel szemben (*Abusalah, Oakes és Tait, 2009.*).

8. Összegzés

Mint láthattuk, rengeteg féle megközelítés létezik a többnyelvű információkeresés problémájának megoldására. A munkámban bemutattam milyen kérdések vetődnek fel a keresés során, azokat hogyan lehet, ill. lehetne megoldani, valamint hogy az egyes módszerek mennyire alkalmasak erre a feladatra. A legfontosabb eljárások mellett több kiegészítő, illetve ritkábban használt megközelítésről is beszéltem.

Megfigyelhető, hogy bármennyire is sikeresek ezek a kísérletek, az egynyelvű keresés hatékonyságát csak különleges esetekben, mint például forrásgazdag környezetben, érik el, esetleg múlják felül. Az is elmondható, hogy minél jobban hozzászoknak az emberek rendszerekhez, annál jobban tudják kihasználni a benne rejlő lehetőségeket. Remélhetőleg tovább folytatódnak a kísérletek és egyre jobb eredményeket produkálnak az újonnan kifejlesztett megoldások, tekintve hogy bőven van még hova fejlődni a területnek, különösen az egynyelvű kereséshez képest, ahol már az apró javulás elérése is sikernek számít, hála a fejlett keresőrendszerek magas színvonalának (*Nie, 2010.*).

Bár a dolgozat nem tér ki rá, de a többnyelvű információkeresés nem csak írásban működhet, hanem szóban is. Ez a kérdés mai felgyorsult világban rendkívül fontos szerepet kaphat, hiszen az információ özönvízszerűen lepi el az embereket. Azonban ez egy még nagyobb feladat elé állítja a kutatókat, hisz a lehetséges problémák – nyelvváltozatok és nyelvjárások, kiejtési gondok, akcentusok, stb. –, melyek felmerülnek a beszédben, csak meghatványozzák a nehézségeket, amikkel szembe kell nézniük.

Irodalomjegyzék

Abusalah, M., Oakes, M., Tait, J. (2009): Cross language information retrieval using multilingual ontology translation and query expansion base. *Polibits*, **2**. 40. sz. 13-16.

Ballesteros L., Croft, W. B. (1996): Dictionary-based methods for cross- lingual information retrieval. In: *Proceedings of the 7th international DEXA conference on database and expert systems applications*. Springer-Verlag, London. 791-801.

Ballesteros, L., Croft, W. B. (1997): Phrasal translation and query expansion techniques for cross-language information retrieval. In: *SIGIR '97 proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York. 84-91.

Ballesteros, L., Croft, W. B. (1998): Resolving ambiguity for cross-language retrieval. In: *SIGIR 98' proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York. 64-71.

Braschler, M., Peters, C., Schäuble, P. (1999): Cross-language information retrieval (CLIR) track overview. In: *Trec-8 report*. NIST, Washington. 25-34.

Broglio, J., Callan, J. P., Croft, W. B. (1994): INQUERY system overview. In: *Proceedings of TIPSTER text program*. ACL, Stroudsburg. 47–67.

Brown, P., Lai, J., Mercer, R. (1991): Aligning sentences in parallel corpora. In: *Proceedings of ACL conference*. ACL, Stroudsburg. 169-176.

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. (1993): The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. **19**. 263-312.

Buckley, C., Mitra, M., Walz, J., Cardie, C. (2000): Using clustering and superconcepts within SMART: TREC 6. *Information processing and management*, **36**. 1. sz. 109-131.

Cao, G., Gao, J., Nie, J.-Y., Bai, J. (2007): Extending query translation to cross-language query expansion with Markov chain models. In: *Proceedings of CIKM conference*. ACM, New York. 351-360.

Franz, M., McCarley, J. S., Roukos, S. (1999): Ad hoc and multilingual information retrieval at IBM. In: *Proceedings of TREC conference*. NIST, Gaithersburg. 157-168.

Gale, W. A., Church, K. W. (1993): A program for aligning sentences in bilingual corpora. *Computational linguistics*, **19**. 3. sz. 75-102.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., Huang, C. (2001): Improving query translation for cross-language information retrieval using statistical models. In: *Proceedings of SIGIR conference*. ACM, New York. 96-104.

Gollins, T., Sanderson, M. (2001): Improving cross language retrieval with triangulated translation. In: *Proceedings of SIGIR conference*. ACM, New York. 90-95.

Grefenstette, G. (1999): The World Wide Web as a resource for example-based machine translation tasks. In: *Proceedings of ASLIB translating and the computer 21 conference*. ASLIB, London.

Hedlund, T., Järvelin, K., Keskustalo, H., Prikola, A. (2001): Dictionary-based cross language information retrieval: problems, methods, and research findings. *Information retrieval*, **4**. 3-4. sz. 209-230.

Hiemstra, D., De Jong, F. (1999): Cross-language information retrieval in twenty-one: using one, some or all possible translations. In: *Language technology in multimedia information retrieval twente workshop on language technology 14*. University of Twente, Twente. 19-26.

Hollink, V., Kamps, J., Monz, C., Rijke, M. de (2004): Monolingual Retrieval for European Languages. In: *Information Retrieval*. **7**. 1-2. sz. 33-52.

Hull, D. A., Grefenstette, G. (1996): Querying across languages – A dictionary-based approach to multilingual information retrieval. In: *Proceedings of the 19th ACM SIGIR conference*. ACM, New York. 49-57.

Jang, M.-G., Myaeng, S. H., Park, S. Y. (1999): Using mutual information to resolve query translation ambiguities and query term weighting. In: *Proceedings of ACL conference*. ACL, Stroudsburg. 223-229.

Kay, M., Röscheisen, M. (1993): Text-Translation Alignment. *Computational linguistics*, **19**. 1. sz. 121-142.

Kirchner, J. S., Nuger, J., Zhang, Y. (2009): An extensible crosslinguistic readability framework. In: *BUCC '09 proceedings of the 2nd workshop on building and using comparable corpora: from parallel to non-parallel corpora*. Association for Computational Linguistics, Szingapúr. 11-18.

Lehtokangas, R., Airio, E., Järvelin, K. (2004): Transitive dictionary translation challenges direct dictionary translation in CLIR. *Information processing and management*, **40**. 6. sz. 973-988.

Levow, G.-A., Oard, D. W., Resnik, P. (2005): Dictionary-based techniques for crosslanguage information retrieval. *Information processing and management*, **41**. 3. sz. 523-547.

Liu, Y., Jin, R., Chai, J. Y. (2005): A maximum coherence model for dictionary-based crosslanguage information retrieval. In: *Proceedings of SIGIR conference*. ACM, New York. 536-543.

Lu, W.-H., Chien, L. F., Lee, H.-J. (2004): Anchor text mining for translation of Web queries – a transitive translation approach. *ACM transactions on information systems*, **22**. 2. sz. 242-269.

McCarley, J. S. (1999): Should we translate the documents or the queries in cross-language information retrieval. In: *Proceedings of ACL conference*. ACL, Stroudsburg. 208–214.

McNamee, P., Mayfield, J. (2002): Comparing cross-language query expansion techniques by degrading translation resources. In: *Proceedings of SIGIR conference*. ACM, New York. 159-166.

Metzler, D., Croft, W. B. (2005): A Markov random field model for term dependencies. In: *Proceedings of SIGIR conference*. ACM, New York. 472-479.

Nie, J.-Y. (2010): *Cross-language information retrieval*. Morgan & Claypool, San Rafael.

Nie, J.-Y. (2003): Cross-language information retrieval. In: *IEEE Computational Intelligence Bulletin*. 2. 1. sz. 19-24.

Nie, J.-Y., Simard, M., Isabelle, P., Durand, R. (1999): Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York. 74-81.

Peters, C. (1997): Across languages, across cultures - Issues in multilinguality and digital libraries. *D-Lib magazine*, 3. 5. sz.

Peters, C., Sheridan, P. (2001): Multilingual information access. In: *Lectures of information retrieval*. Springer-Verlag, New York. 51-80.

Pirkola, A. (1998): The effects of query structure and dictionary setups in dictionary-based crosslanguage information retrieval. In: *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York. 55-63.

Prószéky, G. (2005): An Approach to Machine Translation via the Rule-to-Rule Hypothesis. In: Hutchins, J., Kis, B., Prószéky, G. (szerk.): *Proceedings of the 10th conference of the European Association for machine translation (EAMT)*. PPKE IKT, Budapest. 219-225.

- Rozic-Hristovski, A., Humar, I., Hristovski, D. (2002): Developing a multilingual, personalized medical library portal – Use of MyLibrary in Slovenia. *Electronic library and information systems*, **37**. 3. sz. 146-157.
- Salton, G. (1970): Automatic processing of foreign language documents. *Journal of the american society for information science*, **21**. 3. sz. 187-194.
- Sheridan, P., Ballerini, J. P. (1996): Experiments in multilingual information retrieval using the SPIDER system. In: *Proceedings of SIGIR conference*. ACM, New York. 58-65.
- Simard, M., Foster, G., Isabelle, P. (1992): Using cognates to align sentences in bilingual corpora. In: *Proceedings of the fourth international theoretical and methodological issues in mechanical translation*. Montreal. 67-81.
- Soergel, D. (1997): Multilingual thesauri in cross-language text and speech retrieval. In: *Working notes of AAAI spring symposium on cross-language text and speech retrieval*. University of Maryland, College Park. 164-170.
- Sparck Jones, K. (1998): Information retrieval: how far will *really* simple methods take you?. In: Hiemstra, D., De Jong, F., Netter, K. (szerk.): *Proceedings of the 14th Twente Workshop on language technology*. University of Twente, Twente. 71-78.
- Talvensaari, T. (2008): *Comparable corpora in cross-language information retrieval - Academic dissertation*. University of Tampere, Tampere.
- Ungváry, R., Vajda, E. (2002): *Könyvtári Információkeresés*. Typotex, Budapest.
- Uszkoreit, H. (1998): Cross-Language Information Retrieval: From naive concepts to realistic applications. In: Hiemstra, D., De Jong, F., Netter, K. (szerk.): *Proceedings of the 14th Twente Workshop on language technology*. University of Twente, Twente. 1-8.
- Vauquois, B. (1968): A survey of formal grammars and algorithms for recognition and transformation in machine translation. In: *Proceedings of IFIP congress-68*. North-Holland, Amsterdam. 1114–1122.

Vet, P. E. van der, Bakel, B. van (1998): Combining linguistic and knowledge-based engineering for information retrieval and information extraction. In: Hiemstra, D., De Jong, F., Netter, K. (szerk.): *Proceedings of the 14th Twente Workshop on language technology*. University of Twente, Twente. 59-70.

Xu, J., Weischedel, R. (2005): Empirical studies on the impact of lexical resources on CLIR performance. *Information processing and management*, **41**. 3. sz. 475-487.