Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás

Pohl Gábor

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

Kivonat: A minta-alapú gépi fordítás (EBMT) alapfeltétele, hogy forrásnyelvi és ezeknek megfelelő célnyelvi mintamondatok mondatnál kisebb szerkezeti egységeit automatikusan egymáshoz tudjuk rendelni. Cikkünkben egy EBMT alapú angol-magyar fordítómemóriához (MetaMorpho TM) kidolgozott főnévicsoport-szinkronizáló algoritmust, valamint egy magyar főnévi csoportok angol megfelelőik alapján történő meghatározására kifejlesztett módszert mutatunk be. A főnévi csoportok szinkronizálása során módszerünk tövesített szótári keresést alkalmazva, hasonló alakú szavakat (*cognate*), illetve szófaji egyezéseket keresve minden lehetséges főnévicsoport-párhoz kiszámít egy heurisztikus hasonlósági értéket, majd ez alapján dönt az egyes főnévi csoportok egymáshoz rendeléséről. A szintaktikai elemzővel meghatározott angol főnévi csoportok magyar megfelelőinek meghatározására kidolgozott módszerünk magyar szintaktikai elemzőt nem igényel, az angol főnévi csoportok szavait szótár segítségével képezi le a magyar mondat szavaira, majd a lehetséges fedések közül a magyar mondatra legrövidebben illeszkedőt teljes magyar főnévi csoporttá bővíti (a szótárral meg nem feleltetett szavak szófaját is figyelembe véve a bővítés során). Cikkünkben végül az első szinkronizációs eredményeinket is ismertetjük.

1 Bevezetés

A minta-alapú gépi fordítás (EBMT, [10]) alapfeltétele, hogy a rendelkezésünkre álló forrásnyelvi és az ezeknek megfelelő célnyelvi mintamondatok mondatnál kisebb szerkezeti egységeit automatikusan egymáshoz tudjuk rendelni. Ezt a folyamatot, nevezzük szinkronizációnak vagy párhuzamosításnak[25].

Ebben a cikkben a MorphoLogicnál fejlesztett EBMT alapú, főnévi csoportokat és mondatvázakat kezelni képes MetaMorpho TM fordítómemória rendszerben [2] alkalmazott főnévicsoport-szinkronizáló modul [4] fejlesztése során kidolgozott módszereket, illetve az eddig elért új eredményeket mutatjuk be.

A MetaMorpho TM rendszer a hagyományos, karakteralapú hasonlósági keresést alkalmazó fordítómemóriákkal szemben a Hodász Gábor által kidolgozott nyelvi

[25] Angolul *alignment*, amelyen nem csak a folyamatot, hanem annak eredményét is értjük.

hasonlósági mértéket [3] alkalmazva képes a lefordítandó mondathoz; ennek főnévi csoportjaihoz, illetve a mondatból a főnévi csoportokat kiemelve kapott mondatvázhoz hasonló, ismert fordítású mintákat keresni. A megfelelő morfológiai alakok generálásával a rendszer képes a megtalált mondatvázakba az eredetiektől különböző, de ismert fordítású főnévi csoportokat beilleszteni, így a csak teljes mondatokat kezelő fordítómemóriákénál nagyobb fedés (*recall*) érhető el, miközben a pontosság csak akkor csökken, ha a mondatváz vagy a benne található főnévi csoport fordítása a többi mondatrésztől függ. Ugyanakkor a részekből összerakott mondat még az utóbbi estekben is megkönnyítheti a fordító munkáját, hiszen néhány szó vagy szóalak változtatásával feltehetően gyorsabban tud jó fordítást találni, mintha az egész mondatot kellene lefordítania.

A mondatvázból a jövőben a főnévi csoportokon kívül más mondatrészek is kiemelhetők lehetnek majd, így tovább növelve a fedést; ugyanakkor más frázisokat megkötések nélkül kiemelve a mondatvázból a mondatváz fordításának pontossága jelentősen csökkenhet is, ezért a további vizsgálatok elvégzéséig, egyelőre csak főnévi csoportokat kezel a MetaMorpho TM.

Ahhoz, hogy ismert fordítású főnévi csoportokat tudjunk keresni a memóriában, a lefordított mondatpárok forrásnyelvi és célnyelvi főnévi csoportjait egymáshoz kell rendelni. Ha ezt a szinkronizálási feladatot a fordítóra bíznánk, az – egy csak teljes mondatokat tároló fordítómemória egyszerűségéhez képest – túl sok plusz munkát követelne tőle. A plusz munkára fordított idő pedig lehet, hogy hosszú távon se térülne meg a több keresési találat által megtakarított fordítási idővel. Ezért a főnévicsoport-szinkronizáció automatizálása mellett döntöttünk. Az automatikus főnévicsoport-szinkronizáció, teljes pontosságot biztosító módszer hiányában hibaforrásként jelenik meg a MetaMorpho TM rendszerben, cikkünkben azonban egy olyan egyszerű, szótáralapú módszert fogunk bemutatni, amelyről első eredményeink alapján azt állíthatjuk, hogy megfelelő pontosságot biztosít, abban az esetben is, ha a magyar mondatok főnévi csoportjait szintaktikai elemző nélkül, az angol megfelelőik ismeretében, egyszerű heurisztikával határozzuk meg.

2 Automatikus főnévicsoport-szinkronizáció

Ebben a szakaszban pontosítjuk a főnévi csoportok automatikus szinkronizációjának részben már ismertetett fogalmát; rámutatunk az automatikus megvalósítás eredendő nehézségeire; bemutatjuk a korábbi hasonló módszerek fő jellemzőit, illetve új módszerünk kidolgozásának okait; végül pedig részletesen ismertetjük az újonnan kidolgozott módszerünket.

2.1 Az automatikus főnévicsoport-szinkronizáció feladata

Az automatikus főnévicsoport-szinkronizáció során egymás fordításának tekinthető mondatpárok főnévi csoportjait algoritmikus módszerekkel rendeljük egymáshoz. Az összerendelés során egyes főnévi csoportok pár nélkül maradhatnak. A pár nélkül maradás oka lehet a nyelvek szintaktikai különbözősége (1. példa), állandósultnak tekinthető lexikai különbség (2. példa), vagy a fordító döntése, hogy a természetesebb hangzás érdekében átfogalmazza a mondatot (3. példa). Utóbbi esetben a mondatpár

főnévi csoportjai között lehetnek olyanok, amelyek csak részben tekinthetők egymás fordításának, az ilyen párok szinkronizálása nem lehetséges. (Az automatikus módszer hibájának kell tekinteni, ha mégis rögzít egy csak részben megfeleltethető párt.)

[I] have read [his new book on bread baking].
Elolvastam [a kenyérsütésről szóló új könyvét]. (1. példa)

1. példa: Az angol *I* személyes névmáshoz nem található a magyar fordításban neki megfeleltethető főnévi csoport. (A példákban a maximális méretű főnévi csoportokat szögletes zárójelek határolják.)

[Lolek] had [a huge breakfast].
[Lolek] jól megreggelizett. (2. példa)

2. példa: A *have a huge breakfast* angol kifejezésen belüli főnévi csoportnak nincs párja a magyar fordításban.

[Csabi] ate [ice-cream].
[Csabi] [fagyit] evett. (3. példa)
[Csabi] fagyizott.

3. példa: Ha a fordító az alsó sorban látható fordítást választja, akkor az *ice-cream–fagyi* pár nem határozható meg.

2.2 A főnévi csoportok azonosításának nehézségei

A főnévi csoportok szinkronizálásának veszélyeire már az előző egyszerű példák is rámutattak, nehézségek azonban már a szinkronizálás előtt, a főnévi csoportok automatikus azonosításakor is jelentkeznek. A főnévi csoportok határai sok esetben bizonytalanok (4. és 5. példa), így automatikus meghatározásuk nehéz.

[Ez a királypingvin] éhes.
[Ez] [a királypingvin], [az] pedig [a császárpingvin]. (4. példa)

4. példa: Az *ez* szó a második mondatban külön főnévi csoport.

I saw [the man] in [the garden].
I know [the man in the garden]. (5. példa)

5. példa: Az első mondatban az *in the garden* szabad mondatbővítmény, míg a második esetben a főnévi csoport módosítója.

Ha az alkalmazott szintaktikai elemző különbözőképp határozza meg a főnévi csoportok határait a tárolt mondatpárok forrás- és fordításoldalán, a helyes szinkronizáció elérése akadályba ütközik. Felmerül a kérdés, hogy ilyen esetekben miért nem választjuk ki a helyes elemzést a fordítás alapján. Ha az egyik szintaktikai elemzőt megbízhatónak tekintenénk, akkor a szövegpár másik oldalán korlátozott

mértékben lehetőség lehetne az elemzés egyértelműsítésére, azonban ez nem kívánt mellékhatásokkal is járhatna, hiszen, mint a 2. és 3. példák mutatják, egyáltalán nem biztos, hogy egy adott főnévi csoport fordítása módosítások nélkül jelenik meg a fordításban.

2.3 Korábbi módszerek

Egyszerű, rövid főnévi csoportok („*noun phrase chunk*") szinkronizálására Julian Kupiec 1993-ban ismertetett egy korpuszalapú módszert [6], amely jó ötleteken alapult, azonban mai szemmel nézve viszonylag alacsony pontossága, illetve a lassú (offline) feldolgozás szükségessége nem felelt meg a MetaMorpho TM rendszerbe való integrálás követelményeinek.

A főnévi csoportok szinkronizálásához hasonló fordításkeresési problémákkal foglalkoznak a statisztikai gépi fordító (SMT) rendszerek szinkronizáló algoritmusai, ugyanakkor ezek a módszerek a tanulási fázisban Kupiec módszerénél is nagyobb párhuzamos korpuszt és komoly számítási kapacitást igényelnek, így a MetaMorpho TM rendszerben kezdetben nem kívántuk alkalmazni őket. A statisztikai módszerek másik jelentős gondja, hogy szintaktikai ismeretek híján a főnévi csoportok határait nem tudják pontosan meghatározni. Utóbbira azonban külön szintaktikai elemző (esetleg sekély elemező) alkalmazásával lehetőség lenne.

A mondatnál kisebb egységek szinkronizációs módszereinek másik fő csoportját az elemzésifa-szinkronizáló (*parse-tree alignment*) módszerek alkotják, amelyekkel a közelmúltban bíztató eredményeket értek el [1], de sajnos csak nagyon hasonló elemzési fák esetén, így az angol-magyar nyelvpár kezelésére más módszert kellett kidolgoznunk.

2.4 Angol-magyar szótáralapú főnévicsoport-szinkronizáció

Új főnévicsoport-szinkronizáló módszerünk kidolgozásakor célunk a sebesség és a pontosság (*precision*) maximalizálása volt. Nagy sebességre azért van szükség, mert a tárolt párok főnévi csoportjainak fordításait jogosan várhatja a felhasználó akár már a következő mondat fordításakor is, hiszen a hagyományos fordítómemóriák is gyorsan tárolják, és azonnal elérhetővé is teszik a fordításokat. A pontosság igénye a bevezetés után talán nem szorul részletes magyarázatra, a hibás párok később hibás fordítási ajánlatokhoz vezetnek, ezért a pontosság növelése akár a fedés (*recall*) csökkenése árán is elfogadható.

A kellő gyorsaság elérése érdekében a korpuszalapú módszerek lassú (*offline*) elemzési lépései helyett gyors, tövesített szótári keresést, hasonló alakú szó (*cognate*) [9] keresést és szófaji egyezés keresést alkalmazó megoldás mellett döntöttünk.

A szinkronizáció során minden lehetséges főnévicsoport-párhoz kiszámítunk egy heurisztikus hasonlósági értéket, majd azokat a párokat jelöljük meg összetartozóként, ahol a hasonlósági érték egy küszöbértéknél nagyobb, és a pár mindkét főnévi csoportja a párbeli társára hasonlít leginkább. Utóbbi kitétel azt jelenti, hogy a lehetséges jó párok közül a legjobbat választjuk. Választásra a gyakorlatban csak akkor kényszerülünk, ha egy mondatban legalább két nagyon hasonló (vagy azonos, azaz ismétlődő) főnévi csoportot találunk.

2.4.1 Főnévi csoportok hasonlósága

A különböző nyelvű főnévi csoportok hasonlóságának vizsgálatakor célunk egyetlen, mostantól hasonlósági értéknek nevezett skalár meghatározása. A hasonlósági vizsgálat során az összehasonlított két főnévi csoport tokenjeit (~szavait) egymás után többféle módon is megpróbáljuk egymásnak megfeleltetni, majd az egyes módszerek által lefedett tokenek számából számítjuk ki (heurisztikusan) a hasonlósági értéket az 1. képletben meghatározott módon.

Először tövesített szótári keresést alkalmazunk: a forrásnyelvi főnévi csoport szavainak lehetséges töveit keressük egy speciális, tövesített indexet és találatlistát tartalmazó szótárban, majd a találatok közül csak azokat hagyjuk meg, amelyek a forrásoldalra illeszthetők és fordításuk minden szavának legalább egy lehetséges töve megtalálható a fordításbeli főnévi csoportban. A tövesített index egy keresett tőhöz tetszőleges számú, a forrásoldalon egyszavas kifejezésre tárol mutatót, viszont csak maximalizált számú többszavas kifejezéspárt tesz megtalálhatóvá. Utóbbi azt eredményezi, hogy a gyakori szavak kifejezései csak a többi, kisebb gyakoriságú kifejezésalkotó szót keresve találhatók meg, viszont a kifejezéskeresés tere stopword lista nélkül is jelentősen csökken. A szótári keresés után a forrásoldali főnévi csoport minden tokenjéhez hozzárendeljük a környezetére leghosszabban illeszthető találatokat, ezzel szűrve az elfedett rövidebb kifejezéseket (6. példa).

This is a {hard disk drive}.
In the first {drive} slot there is a {hard disk drive}. (6. példa)

6. példa: Az első mondatban a *hard disk drive* kifejezés elfedi a lehetséges *hard disk, disk drive, disk, drive, hard* találatokat. A *hard disk drive–merevlemez meghajtó*, illetve *drive–meghajtó* szótári találatokat feltételezve az utóbbi pár felvétele hibás lenne. A második példában viszont a *drive* – független előfordulása miatt – mégis szerepelhet a szótári találtatok között, ha a fordításban is megtalálható.

A szótári megfeleltetés után, a főnévicsoport-pár le nem fedett szavai között hasonló alakúakat (*cognate*, pl. az angol *parliament* és a magyar *parlament* szavak) keresünk a Simard és társai által kidolgozotthoz [9] nagyon hasonló algoritmust alkalmazva. Megvalósításunkban két szót akkor tekintünk hasonló alakúnak, ha egy karakternél hosszabbak, tartalmaznak legalább egy nagybetűt, számot vagy valami más speciális karaktert, és legfeljebb az ötödik karaktertől különböznek. (Az ismertetettnél kevésbé hatékonyan számítható, de nagyobb fedésű algoritmus alkotható a szavak közötti Levenshtein-távolság mérésével.)

A korábban le nem fedett szavakat ezután szófajaik alapján próbáljuk egymáshoz rendelni. Ha egy szóhoz több lehetséges szófajt is rendelt a morfológiai elemző, bármelyikkel való egyezést elfogadunk.

A lefedetlen szavak közül a pusztán grammatikai funkciót betöltőket egy kis büntetőpontszámot alkalmazva kiemeljük, így téve lehetővé a csak a grammatikai funkciót betöltő szavaikban különböző rövid főnévicsoport-párok egymáshoz rendelését (7. példa).

Where is [**my** book]?
Hol [**a** könyvem]? (7. példa)

7. példa: A *my book* megfelelője az *a könyvem* főnévi csoport, ugyanakkor szinkronizálásukkor gondot okozhatna, hogy a 4 szó közül 2 nem feleltethető meg egymásnak, ha nem tekintünk el a pusztán grammatikai funkciót hordozó, egymásnak meg nem feleltethető szavaktól.

Az előzőekben ismertetett illesztések után a *h* hasonlósági értéket az alábbi 1. képlet alapján számítjuk ki.

$$h = \frac{a \cdot Dict + b \cdot Cogn + c \cdot POS - d \cdot GF}{T - GF} \qquad \text{(1. képlet)}$$

1. képlet: A hasonlósági érték számításának módja. A képletben *Dict* a szótárral megfeleltetett tokenek száma, *Cogn* a hasonló szavakat keresve lefedett tokene száma, *POS* a szófaji illesztéssel lefedett tokenek száma, *GF* a le nem fedett pusztán grammatikai funkciójú tokenek száma, *T* pedig a két főnévi csoport tokenszámának összege. Az a=1, b=0.9, c=0.3, d=0.1 konstans együtthatók, empirikusan meghatározott értékekkel.

Az 1. képlet kísérletezéssel, de csak kevés mintán meghatározott konstans együtthatóit (vagy akár magát a képletet), a jövőben a főnévi csoportok közötti kapcsolatokat is tartalmazó párhuzamos korpusz vizsgálatával kívánjuk finomítani.

3 Magyar főnévi csoportok meghatározása angol megfelelőik alapján

A MetaMorpho TM rendszerben a tárolt mondatpárok angol oldalán a MetaMorpho elemzőt és a hozzá készített angol (valójában angol-magyar) nyelvtant [11] használjuk a főnévi csoportok meghatározására. A magyar oldal megfelelő pontosságú elemzésre azonban még nem alkalmas az elemzőhöz fejlesztett magyar-angol nyelvtan [8], így megpróbáltuk az angol elemzővel automatikusan meghatározott főnévi csoportokhoz rendelhető magyar főnévi csoportokat az angol főnévi csoportok szavait és kifejezéseit a magyar szövegre leképezve meghatározni.

Heurisztikus megoldásunkban a 2.4.1. pontban bemutatott módszerekkel minden angol főnévi csoport szavaihoz szótári egyezéseket és hasonló alakú szavakat keresünk. A keresés során különbség a 2.4.1. pontban ismertetett módszerhez képest, hogy a grammatikai funkciót betöltő szavakat nem keressük a szótárban, mivel ezek fordítása a magyar mondatban a keresett főnévi csoporttól függetlenül bárhol előfordulhat. Azokat a szavakat tekintjük grammatikai funkciót betöltőnek, amelyek morfológiai elemzéssel meghatározott lehetséges szófajai között csak néhány, előre meghatározott szófaj (névmás, névelő stb.) szerepel.

Mivel egy szó akár többször is előfordulhat a mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a magyar mondatra. Természetesen a találatok között más szavakat is tartalmazhat a kije-

lölt illeszkedés. A legrövidebb illeszkedés kiszámítása költséges művelet, érdemes a fedéshossz korlátozásával redukálni a keresési teret.[26]

Az illeszkedést ezek után egyszerű szabályok és az angol főnévi csoport le nem fedett szavainak figyelembe vételével teljes magyar főnévi csoporttá bővítjük. A lényegesebb szabályokat a következőképp foglalhatjuk össze. Először a találatok közötti szavak szófaját próbáljuk az angol főnévi csoport meg nem talált szavainak szófajával egyeztetni, majd ha pár nélküli melléknév vagy főnév szerepel az angol főnévi csoportban, akkor baloldalról próbáljuk bővíteni a magyar főnévi csoportot. A bővítés során az angol főnévi csoport pár nélküli főneveinél, illetve mellékneveinél maximum eggyel többet engedünk meg, illetve nem folytatjuk a bővítést, ha igéhez vagy más a főnévi csoportba nem illő szóhoz, írásjelhez érünk. Jobbra csak akkor bővítjük a főnévi csoportot, ha a baloldali bővítési kísérlet után is maradt páratlan főnév az angol főnévi csoportban. A főnévi csoportot mindig kibővítjük a tőle balra található névelővel. (A módszer a megvalósításban kicsit bonyolultabb, mivel szófaji egyértelműsítés hiányában az egyes szavak szófajai esetében több lehetőség közül kell választanunk.)

A módszer egyszerűségéből és a szótári találatok bizonytalanságából adódóan néha jelentős hibákat ejt, ezeket azonban a 2.4.1. pontban ismertetett módszerrel könnyen szűrni lehet: ha az angol párja alapján meghatározott magyar főnévi csoport nem hasonló eléggé angol párjára, akkor a párt nem rögzítjük.

A módszer jelenlegi formájában egyesével, egymástól függetlenül választja ki az angol főnévi csoportokhoz rendelt párokat, így hibázás esetén akár átfedő párok is kialakulhatnak, bár átfedés esetén nagyon kicsi az esélye annak, hogy mindkét pár elérje a szükséges hasonlósági pontszámot. Ha ez mégis megtörténne, akkor a jelenlegi megoldásban mindkét párt elvetjük. A jövőben meg fogjuk vizsgálni, hogy jobb megoldás lenne-e, ha – balról jobbra haladva – a szükséges hasonlósági pontszámot megszerző főnévi csoportok által lefedett szavakat foglaltnak jelölnénk, és nem használnánk őket más főnévi csoportban; illetve megpróbálunk módszert kidolgozni arra, hogy még a főnévi csoportok bővítése előtt feltérképezzük viszonyaikat a magyar mondatban.

4 Eredmények

Első kísérleteinket az informatikai témájú szövegeket tartalmazó SZAK-korpusz [7] 40 viszonylag hosszú mondatpárt (átlagosan 23 szó/mondat) tartalmazó kis részletén végeztük. A kísérlethez viszonylag kis méretű, 116 000 szó- és kifejezéspárt tartalmazó szótárt használtunk.

Az automatikusan meghatározott angol főnévi csoportok 56 százalékának volt csak meghatározható magyar fordítása. (Az angol mondatok alanya gyakran személyes névmás volt, a fordító néha igei szerkezetre cserélte a főnévi csoportot, néhány esetben pedig az angol elemző hibázott.) Módszerünk pontossága 84 % volt, azaz a fordítómemóriába felvett párok kevesebb, mint 1/6 része hibás.

---

[26] Megjegyzés: A paraméteres bonyolultságelmélet rámutat arra, hogy a bonyolult (NP-teljes) feladatok között több olyan is van, amelyeknél bizonyos paramétereik rögzítésével redukálható a keresési tér, így könnyű (polinomiális) feladattá vezethetők vissza (természetesen csak rögzített paraméterek mellett).

Az eredetitől jelentősen különbözően fordított mondatokat elhagyva azt tapasztaltuk, hogy a pontosság 91 százalékra nőtt, azaz ha a fordítómemória felhasználója a felhasználói felületen (egyetlen kattintással, vagy billentyűkombinációval) megjelölhetné a részekre nem bontható fordításokat, akkor a pontosság jelentősen növelhető lenne. A párosítható angol főnévi csoportokhoz mérve 65 százalékos fedést (*recall*) sikerült elérni, ami a szótár bővítésével, reményeink szerint még növelhető. (A 65%-os fedés azt jelenti, hogy az angol főnévi csoportok kicsit több, mint 1/3 részéhez rendelt fordítást a módszerünk.)

A 2.4. és 3. pontokban ismertetett módszerek sebessége megfelelő, együttes futásidejük a próbák során legfeljebb néhány ezredmásodperc volt egy átlagos számítógépen (igaz a módszereket a hatékonyságra ügyelve implementáltuk), ez az angol főnévi csoportok megállapítására használt szintaktikai elemzés idejéhez képest elhanyagolható.

5 További tervek

Az eddiginél nagyobb mintán végzett mérésekhez kézzel címkézett tesztanyag, főnévicsoport-szinten párhuzamosított korpusz építése szükséges, így ez rövidtávú terveink között szerepel. A tesztanyag méretének növekedtével annak egy részét az 1. képlet konstans paramétereinek behangolására szeretnénk fordítani.

A szinkronizáló módszer önálló mérése mellett azt is szeretnénk megvizsgálni, hogy egyes hibái hogyan hatnak a teljes MetaMorpho TM rendszerre. A teljes rendszer tesztelését a Hodász Gábor által kidolgozott módszerekkel [5] végezzük. Az eredményeket elemezve kell majd módszert adnunk a memóriába került hibás párok szűrésére, illetve az esetlegesen kritikus hatású hibák elkerülésére.

A hatékonysági mérésekkel párhuzamosan, korpuszalapú, statisztikai szótárbővítő módszereket alkalmazva gyarapítjuk majd a szinkronizációhoz használt szótárat; illetve a magyar főnévi csoportok meghatározására kidolgozott, 3. pontban ismertetett módszer helyett a MetaMorpho elemzőhöz készülő magyar nyelvtant is ki fogjuk próbálni.

Hivatkozások

1. Groves, D.; Hearne, M.; Way, A.: Robust Sub-Sentential Alignment of Phrase-Structure Trees. COLING '04, Geneva, Switzerland, 2004
2. Grőbler Tamás, Hodász Gábor, Kis Balázs: MetaMorpho TM: A Rule-Based Translation Corpus. International Conference on Language Resources and Evaluation, Lisszabon, 2004.
3. Hodász Gábor: Nyelvi hasonlóságon alapuló intelligens keresés fordítómemóriában. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged
4. Hodász Gábor, Pohl Gábor: MetaMorpho TM: a linguistically enriched translation memory. In: International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan, ISBN 954-90906-9-8), Borovets, Bulgaria, 24 Sept. 2005.
5. Hodász Gábor: Fordítómemóriák és mintaalapú fordító rendszerek kiértékelésének módszerei. ugyanebben a kötetben

6. Julian Kupiec: An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora. In: Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 17-22, 1993
7. Kis Ádám; Kis Balázs: A Prescriptive Corpus-based Technical Dictionary. In: Papers in Computational Lexicography: Proceedings of COMPLEX 2003. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2003.
8. Merényi Csaba: A MetaMorpho magyar-angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan. ugyanebben a kötetben
9. Simard, M., Foster, G. & Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation, (TMI92), Montreal, pp. 67-81, 1992
10. Somers, H.: An Overview of EBMT. In M. Carl. and A. Way. (eds.) Recent Advances in Example-based Machine Translation, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3-57. 2003.
11. Tihanyi, L.: A MetaMorpho projekt 2004-ben. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2004.