Ph.D-értekezés tézisei

Nagy Ágoston

Terminológiakivonatolás francia nyelvű szabadalmi leírásokból szabály alapú és statisztikai módszerek segítségével

Témavezetők:

Dr. Váradi Tamás
tudományos főmunkatárs

Dr. Gécseg Zsuzsanna
egyetemi docens

Szegedi Tudományegyetem
Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola
Francia nyelvészet alprogram

2012

A kutatás témája

Az automatikus terminológiakivonatolás (a továbbiakban TE) a számítógépes nyelvészet egy alterülete, melynek célja olyan alkalmazás kidolgozása, amely adott írott, nyers szövegből automatikusan kinyeri annak terminusjelöltjeit (Jacquemin és Bourrigault 2003). A TE egyik alkalmazási célja az automatikus szövegindexelés, amely során egy megadott szöveges fájl rá jellemző olyan fontosabb kifejezéseit kivonatoljuk, amelyek a későbbiekben elősegíthetik azok kategorizálását vagy a közöttük történő keresést (Enguehard 2005). A TE egy másik felhasználási területe a fordítói munka elősegítése: az előre kinyert terminusok idegen megfelelőjükkel történő összepárosítása a fordítási munka előtt igencsak hasznos. Szintén terminológiakivonatoló eszközökre támaszkodnak gépi fordítást megvalósító alkalmazások (pl. Vasconcellos 2001), illetve információkinyerő eszközök (pl. Ahmad 2001).

Mind a hazai, mind a nemzetközi szakirodalom elsősorban a főnévi terminusokra koncentrál, mert ezek rendelkeznek a legösszetettebb szerkezettel, ezért a disszertációban is a főnévi terminusok kinyerésével foglalkoztunk. A disszertációhoz készült egy számítógépes alkalmazás, amely ezen típusú terminusokat francia nyelvű szabadalmak leírásaiból nyeri ki. A terminusok kinyeréséhez használunk szabály alapú, azaz nyelvészeti módszereket, illetve statisztikai alapú eljárásokat.

A kutatás elméleti háttere

A klasszikus terminusdefiníció megalkotása Wüster (1976, 1981) nevéhez köthető. Ez az a nézet, amit a terminológiai közösség elismer és elfogad, amikor a terminus meghatározására kerül sor (Petit 2001). E szerint egy lexikai egység terminusi mivolta a következő három feltételhez köthető: (1) a terminus kapcsolódik egy (és csakis egy) fogalomhoz, (2) megnevezi ezt a fogalmat és (3) valamilyen szakterülethez köthető.

A szabadalmak olyan szövegek, amelyekben a szakszövegekre jellemző jegyek – mint szakszavak, személytelen szerkezetek használata (pl. Cabré 1999) – jelentős mértékben fordulnak elő, így a TE számára ez a típusú korpusz ideális. A leírás a szabadalmak legterjedelmesebb része, célja a szabadalom tömör, pontos ismertetése, ezért ezekben a terminusok gyakran ismétlődnek, mert Justeson és Katz (1995) szerint a terminusok egyik ismertetőjegye az, hogy ritkán helyettesíthetőek más terminusokkal vagy rövidebb szóalakokkal. Mindezek és Cabré és mtsai (2001) alapján e szövegeken a statisztikai módszerek jól alkalmazhatók.

Cabré és mtsai (2001), valamint Sauron (2002) szerint a TE-alkalmazások menetének főbb lépései a következők: (1) terminusjelöltek kinyerése, amely egy adott korpuszból történik, (2) ezen terminusjelöltek szűrése, végül (3) a terminusjelöltek validálása. Mind a terminusjelöltek kinyerése, mind azok szűrése történhet szabály alapú módszerekkel (előre megadott mintákkal) vagy statisztikai módszerekkel. A legtöbb TE-alkalmazás hibrid módszert alkalmaz, mivel a szabály alapú kivonatolók túl nagy zajt (tehát a kivonatolt terminusjelöltek száma magasabb, mint a valós terminusoké), a statisztikai alapúak pedig túl nagy csendet okoznak (a terminusjelöltek listája sok terminust nem tartalmaz). A szabály alapú módszer esetében a terminusok a belső morfoszintaktikai szerkezetük segítségével kivonatolhatók, tehát terminusokra jellemző összetételeket kell keresnünk, például két egymás után álló főnevet. A hibrid módszerek közül az előnyben részesített kombináció a statisztikai módszerrel történő terminuskinyerés és a szabály alapú szűrés (pl. Ha és mtsai 2008).

A TE-alkalmazások esetében a statisztikai módszerek három különböző csoportra oszthatók. Az első kategóriába tartoznak a *termhood*-értékek, amelyek azt mutatják meg, hogy az adott terminus mennyire kötődik a szaknyelvhez, illetve a köznyelvhez. Ezen érték kiszámításához szükség van az adott jelölt köznyelvi előfordulási arányára is. A második érték a *unithood*-érték, ami egy terminusjelöltnél azt mutatja meg, hogy annak elemei mennyire tartoznak össze, azaz az adott terminusjelölt tényleg egy egység, egy nagyobb egység része vagy csak annak egy része terminus. A harmadik statisztikai érték azt mutatja meg, hogy a

terminusjelölt környezetében lévő tokenek milyen valószínűséggel előznek meg vagy követnek terminust (Wong és mtsai 2008).

A terminológiakivonatoló hatékonyságát a számítógépes nyelvészetben alkalmazott hagyományos pontosság, fedés és F-értékkel írjuk le (pl. Cabré és mtsai 2001). A pontosság a terminológiakivonatolás esetén azt mutatja meg, hogy a kivonatolt terminusjelölt-listában milyen arányban vannak azok az elemek, amelyek valóban terminusok. A pontosság kiszámítására az alábbi képletet alkalmazzuk:

$$pontosság = \frac{1}{kivonatolt\ terminusok\ száma} \cdot ebből\ valós\ terminusok\ száma$$

Ezzel szemben a fedés azt adja meg, hogy milyen arányban szerepel a terminusjelölt-listában az összes igazi terminus, tehát a csend mértékét adja meg. A fedés kiszámítására az alábbi formulát használjuk:

$$fedés = \frac{kivonatolt\ helyes\ terminusok\ száma}{valós\ terminusok\ száma}$$

A számítógépes nyelvészetben és a matematikai statisztikában gyakran használják még az F-értéket a hatékonyság mérésére. Ez az érték a pontosság és a fedés értékének harmonikus közepe. Így a hagyományos F-érték kiszámításának módja:

$$F\text{-}érték = \frac{2 \cdot pontosság \cdot fedés}{pontosság + fedés}$$

A francia nyelvű főnévi terminusok szerkezete

A szabály alapú terminuskinyeréshez minél pontosabban meg kellett határozni a terminusok különböző lehetséges szerkezeteit, és hogy ezek miben különböznek a hagyományos főnévi csoportoktól, különösen a prepozíciós komplementumok (ide tartozik a főnévi fejet követő prepozíció+főnév szerkezet is az összetett szavak esetében) és a melléknévi adjunktumok tekintetében. Ehhez segítséget nyújtottak korábbi kutatásaink (pl. Nagy 2012a, 2012b) és L'Homme (2004).

A francia nyelv esetében a prepozíciós szintagmák meghatározása nem evidens, mert nem mindig lehet egyértelműen megkülönböztetni a főnévi csoporton belül új entitást bevezető prepozíciós szintagmát (1a) és a főnévi fejhez kapcsolódó és azzal egy összetett főnevet alkotó prepozíciós szintagmát (1b). Az összetett főnevek esetében (1b) a prepozíciót egy determináns nélküli főnév követi, mivel a determináns megléte egy összetett főnévi csoporthoz vezetne, ahol a prepozíciót követő főnévi csoportot egy önálló referenciával rendelkező, beágyazott főnévi csoportnak tekinthetjük (1a).

(1) a. *le* *moulin* *du* *village*
ART-DEF.MASC.SG malom DE+ART-DEF.MASC.SG falu
'a falu malma'

b. *le* *moulin* *à* *vent*
ART-DEF.MASC.SG malom À szél
'szélmalom'

Ezért Bosredon és Tamba (1991) megkülönbözteti a hagyományos prepozíciós szintagmákat (1a) a főnévhez kapcsolódó prepozíció+főnév szekvenciáktól (1b): az előbbieket konstituensnek az utóbbiakat formánsoknak nevezi. Egyértelmű határ azonban nem húzható a kettő közé, ugyanis például vannak olyan összetett főnevek, amelyek olyan prepozíciós szintagmát tartalmaznak, amelyekben a prepozíciót determináns követi (2a).

(2) a. *cancer de la peau*
rák DE ART-DEF.FEM.SG bőr
'bőrrák'
b. ?*cancer de peau*

Egy korábbi tanulmányunkban (Nagy 2009b) bemutattuk, hogy a terminusok esetén a belső determinánsok jelenléte általában 7% az összes terminus között, azonban arra vonatkozólag, hogy ezen 7%-nyi terminus között hány szerepelhet determináns nélkül is, nincs empirikus eredményünk. Ebből adódóan, egy kisebb veszteséggel számolva, nem vettük figyelembe a determinánssal rendelkező terminusokat.

A melléknevek a francia nyelvben állhatnak a főnév előtt és a főnév után is. Cinque (1994) úgy fogalmaz, hogy az újlatin nyelvek, és ezáltal a francia is, az ANA-típusú nyelvek közé tartozik, míg a germán nyelvek inkább AN-típusúak, tehát az utóbbi nyelvekben a melléknév prenominális, az előbbi nyelvekben pedig lehetnek pre- és posztnominálisak is, ez azonban függ a melléknév típusától is.

A négy melléknévcsoport (Riegel és mtsai (2009) szerint ezek a csoportosító, nem csoportosító, relációs melléknevek és sorszámnevek) közül nagyobbrészt azok lehetnek terminusok részei, amelyek a főnévi fej alfaját jelölik. A négy közül ez a tulajdonság a csoportosító és relációs melléknevekre igaz, amelyek csak a főnév után állhatnak: ezt is szemlélteti a (3)-ban szereplő példa, ahol az aláhúzott rész a terminus, a második melléknév a *filaire* 'vezetékes' relációs, az első (*grand* 'nagy') pedig gyakran használt, egyszótagú (tehát prenominális) melléknév, amely kevés eséllyel válhat a terminusok részévé, de előfordulhat azok előmódosítójaként (3).

(3) *un* *grand* *réseau filaire*
ART-INDEF.MASC.SG nagy- MASC.SG hálózat vezetékes
'nagy vezetékes hálózat'

A francia főnévi terminusok nem foglalják magukba az előttük álló determinánst, és a hagyományos főnévi csoportokkal ellentétben nem tartalmazhatnak tagmondatokat sem. Ez következik a sageri (1990) három fő kritériumból: a (1) gazdaságosság (*economy*), (2) pontosság (*precision*), (3) megfelelőség (*appropriateness*), valamint abból, hogy egy terminust mindig ugyanúgy kell használni a wüsteri definíció miatt, ezért a terminusok inkább N'-kategóriájúak. Például a vonatkozói mellékmondatok esetén csak a lehetséges terminuselemek maradhatnak meg, például a szabadalmi korpuszunkból vett (4) esetében csak a *poudre hygroscopique* 'higroszkópos por' maradhat meg terminusként.

(4) *La poudre, qui est hygroscopique, ...*
ART-DEF.FEM.SG por aki/ami/amely van higroszkópos
'A por, ami higroszkópos, ... '

A kutatás célja

Vizsgálatunk során nem a szokásos (statisztikai alapon történő terminuskinyerés, majd szabály alapú szűrés) eljárást alkalmaztuk, hanem ennek fordítottját, ami ritkább a TE során: a terminusjelölt-listát szabály alapú módszerekkel nyertük ki, majd ezt különböző szűrőkkel szűrtük. A terminusjelöltek szabály alapú kinyerésének választásával azt kívántuk bizonyítani, hogy az a szakirodalomban elterjedt nézet (pl. Cabré és mtsai 2001), miszerint a statisztikai módszerekkel történő terminuskinyerés hatékonyabb, nem feltétlenül igaz minden esetben. E választást az is igazolja, hogy (1) a francia nyelvben a főnévi terminusok nagy arányban rendelkeznek olyan belső szerkezettel, ami a köznyelvi egységekre nem jellemző, így a szabály alapú módszerek nagyobb eredményességgel használhatók; valamint az, hogy (2) a francia nyelvre is alkalmazható, legtöbbet idézett terminológiakivonatolókban a kinyerés szabály ala-

pú (például az Acabit (Daille 1994), a Lexter (Bourrigault 1994) és a Fastr (Jacquemin 2001) TE-alkalmazások esetében).

Célunk volt az, hogy bizonyítsuk, a szabály alapú terminuskinyerés magas fedést ér el alacsony pontossággal, amely utóbbin mind a szabály alapú mind a statisztikai szűrési módszerek javítani tudnak. Előfeltételezésünk az volt, hogy a statisztikai módszerek esetében, ha a határértéket magasnak állítjuk be, akkor a pontosság jelentősen megnő, míg a fedés jelentősen csökken; s ha alacsonyabb küszöbértéket választunk, akkor valamelyest lehet az F-értéket növelni.

A disszertáció egyik célja az volt, hogy megmutassuk, milyen mértékben járulnak hozzá a különböző faktorokat figyelembe vevő statisztikai mértékek a szabály alapú kinyerés hatékonyságának növeléséhez, valamint hogy kidolgozzunk egy olyan módszert, amely az alkalmazott három statisztikai mértékből egy összevont értéket képez. Célunk volt még megállapítani, milyen küszöbértéket kell adnunk az összevont statisztikai értéknek ahhoz, hogy a pontosságot jelentősen meg tudjuk növelni a fedés lehető legkisebb csökkenésével.

A terminológiakivonatoló alkalmazás eredményeit ezenkívül más – hasonló feladatot (pl. szövegindexelés) elvégző – programok eredményeivel is összevetettük. Ezen alkalmazások a Fastr (Jacquemin 2001) és a YaTeA (Aubin és Hamon 2006).

A vizsgált korpusz

Az elemzés korpuszát francia nyelvű szabadalmak adták, amelyek javarészt megfelelnek a szaknyelvek jellemzőinek, és terminusok csak ez utóbbi típusú szövegből nyerhetők ki. A szabadalmak több részből állnak, mint például bibliográfiai adatok, absztrakt, leírás vagy igénypontok. Az előbbiek közül a szabadalmi leírást választottuk korpuszként, mert ez a szabadalom legterjengősebb része, célja a találmány pontos és részletes leírása. Ennek következtében igen hosszú és rengeteg szóismétlés jellemzi annak érdekében, hogy még jobban körülírja az adott szabadalmat, így előfeltételezésünk szerint azokra a statisztikai módszerek is jól alkalmazhatók.

Az elemzésünkben két külön szakterülethez tartozó szabadalmi leírást választottunk: az egyik a az informatikai területeket érintő G06F szabadalmi osztály, a másik pedig az emberi szükségletekkel kapcsolatos A23L osztály[1]. Ezen szabadalmi csoportokból tíz-tíz leírást választottunk: a leírások átlagosan körülbelül 4000 szótokennel rendelkeznek, azaz az írásjeleket eltekintve megközelítőleg ennyi szóból állnak. A korpuszban az elemzés előtt kézzel jelöltük be a terminusokat, lehetővé téve ezek és a számítógép által kivonatolt terminusok listájának összevetését. Ez alapján tudja az alkalmazás kiszámítani, hogy a terminusok kinyerésére szolgáló egyes – akár szabály alapú, akár statisztikai – módszerek külön-külön milyen mértékben tudnak hozzájárulni a fedés, illetve a pontosság növekedéséhez vagy csökkenéséhez. A kézi bejelölés folyamán a G06F korpuszban 1752 terminust jelöltünk be, az A23L korpuszban 2086-ot.

A kutatás módszere

A saját alkalmazásban a TE hibrid módszerrel történik: a terminusok kinyerése szabály alapú, azok szűrése pedig szabály alapú és statisztikai. Mivel a szabály alapú terminuskinyeréshez szükség van a benne szereplő szavak szófaji címkéire, valamint a statisztikai módszerekhez ezek szótövére, ezért legelőször egy annotáló alkalmazással (*Connexor*[2]) elvégeztettük ezen műveleteket, így a terminusjelöltek kinyerésekor ezekkel az adatokkal már rendelkeztünk.

A terminusjelölt-lista kinyerése előtt kiszűrtük a szabadalmakban szereplő tulajdonneveket, amire azért volt szükség, mert ezek gyakran előfordulnak szabadalmakban (pl. korábbi

[1] A vizsgálatban a WIPO szervezete által biztosított szabadalmi keresőt, a Patentscope-t használtuk (http://www.wipo.int/pctdb/en/)

[2] http://www.connexor.eu/technology/machinese/demo/

szabadalmakra történő hivatkozásokban), ráadásul főnévnek is jelöli a szófaji címkéző program, így ezek szűrése nélkül sok olyan terminusjelölt került volna a listába, amelyek nem lehetnek terminusok. A tulajdonnevek szűrésére a Javába is beilleszthető *OpenCalais Web Service API*[3] nevű programot használtuk, amely a szövegben a terminuskinyerés számára láthatatlanná tette a személy-, cég-, és helyneveket.

A szabály alapú szűrés során kiszűrjük még a konnektívumokat is, amelyek elsődleges célja Riegel és mtsai (2009) alapján a szöveg strukturálása és kohéziójának fenntartása. Többek között ilyenek a szövegkohéziót biztosító elemek, az *en effet* 'ugyanis' vagy a *par exemple* 'például'. Ezen szókapcsolatokat mindenképpen ki kell zárni a TE során, mert így az előzőekből az *exemple* 'példa' és az *effet* 'hatás' szó nem kerül be hibásan a terminusjelölt-listába.

Azonban a konnektívumok alatt nemcsak a főnevet tartalmazó elemeket értjük, hanem azon mellékneveket, határozószavakat is, amelyek nem lehetnek terminusok részei, ezért ezeket el kellett távolítani a szövegből. Ilyen melléknév például a *suivant* 'következő' vagy a *précédent* 'előző'. Szűrés nélkül például az *acides gras suivants* 'a következő zsírsavak' főnév-melléknév-melléknév kombináció helytelenül kerülne be a terminusjelölt-listába, de az utolsó melléknév eltávolításával (*acides gras* 'zsírsavak') a terminusjelölt már helyes, így mind a fedés (egy elemmel többet ismertünk fel), mind a pontosság nő (egy rossz elemmel kevesebb van a terminusjelölt-listában).

A terminusjelöltek listájának szabály alapú kinyeréséhez először – korábbi tanulmányaink alapján (pl. Nagy 2008, Nagy 2009a, Nagy 2009b) – kigyűjtöttük a terminusokra jellemző szinte összes szintaktikai mintát, például kikerestük a szövegben a legalább két egymást követő főnevet (pl. *accès internet* 'internet-elérés') vagy a főnév és melléknév kombinációkat (pl. *réseau filaire* 'vezetékes hálózat'). A használt minták olyan belső szintaktikai összetételeket is tartalmaznak, amelyek nem csak a terminusokra jellemzőek: ide tartoznak az egyszavas főnevek, amelyek közül például a *terminal* 'terminál' terminus, a *problème* 'probléma' nem. Azon mintázatok, amelyek inkább jellemeznek nem terminusokat, és amelyek felvétele a mintázatok közé inkább rontana az eredményeken, nem kerültek be a listába. Ilyenek voltak a belső determinánst tartalmazó elemek, mint például a *mise à jour du (de+le) site web* 'a webhely fejlesztése', amely esetén mind a *mise à jour* 'frissítés', mind a *site web* 'webhely' külön terminusjelölt lett az utóbbi előtti determináns (*le*) miatt. A korpuszban egy-két terminusjelölt azonban tartalmazott determinánst (pl. a főnév-prepozíció-determináns-főnév mintájú *état des (de+les) lieux* 'tárgyjegyzék'), de ezek száma csekély, így azokat nem vettük fel a lehetséges minták közé.

Ezen minták alapján (pl. főnév-főnév vagy főnév-melléknév) egy determinisztikus és teljes véges állapotú automatát hoztunk létre. Korábbi tanulmányainkban (pl. Nagy 2008) ezt reguláris kifejezésekkel oldottuk meg, de a disszertációban a véges állapotú automaták melletti döntést azok számunkra jobb átláthatósága indokolta. Az állapotok közötti átmenetet kizárólag a szófaji kódokkal címkéztük, és biztosítottuk, hogy az automata csak olyan sorozatot fogadjon el, amely lehetséges terminusminta, és azt, hogy az összes általunk felsorolt mintát felismerje.

A szabály alapú szűrés után alkalmaztuk a statisztikai mértékeket, amelyekkel célunk a pontosság, és ezáltal az F-érték esetleges további növelése volt a fedés értékének lehető legkisebb csökkentésével. A *termhood*-értékek közül az Ahmad és mtsai (1999) által kidolgozott *weirdness*-értéket választottuk, ami egy adott terminusjelölt esetén kiszámolja annak szakszövegbeli és köznyelvbeli előfordulásának arányát. Ha egy kifejezés egy adott szakterületen nagyobb arányban fordul elő, mint egy általános nyelvi korpuszban, akkor az nagyobb valószínűséggel lesz terminus. Az általános nyelvi referenciakorpuszt az interneten – egy online elérhető francia újságban – fellelhető dokumentumok adták, így a terminusjelöltek köznyelvi korpuszban lévő gyakoriságát egy internetes keresőmotor eredményeire alapoztuk: a lekérdezést

[3] http://www.opencalais.com/calaisAPI

az alkalmazás automatikusan hajtotta végre, majd tárolta azt, hogy az hány dokumentumban fordult elő.

A *weirdness*-érték kiszámításának módszere:

$$weirdness(w) = \frac{\frac{f_s(w)}{t_s(w)}}{\frac{f_g(w)}{t_g(w)}}$$

A w a vizsgált szó, $f_s(w)$ a w előfordulási száma a szakszövegben, $f_g(w)$ a w előfordulási száma az általános korpuszban, $t_s(w)$ a szakszöveg összes tokeneinek a száma, míg $t_g(w)$ az általános korpusz tokeneinek száma. Azon szavak esetében, amelyek mind az általános, mind a szakszövegben gyakran és ugyanolyan aránnyal fordulnak elő, ott ezek értéke 1 körüli, míg a terminusoké, amelyek jobban meghatároznak egy szakszöveget, ennél nagyobb.

A *unithood*-értékek közül a Frantzi és Ananiadou (1997), illetve a Maynard és Ananiadou (2000) nevéhez fűződő C-értéket választottuk, amelynek célja azt megállapítani, hogy a jelölt önmagában fordul elő gyakrabban vagy egy nagyobb egység részeként. A C-értéket az alábbi módszerrel számíthatjuk ki:

$$C\text{-}value = \log_2|a| \cdot f(a) \qquad \textit{ha a nem beágyazott}$$

$$C\text{-}value = \log_2|a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \qquad \textit{egyébként}$$

Az a a vizsgált terminusjelölt, $f(a)$ annak előfordulási száma, |a| a terminusjelölt hossza szavakban mérve, T_a azon terminusjelöltek halmaza, amely tartalmazza a-t, $P(T_a)$ pedig azon terminusok száma, amelyek tartalmazzák a-t és hosszabbak is nála.

Ezen érték kiszámításához szükségünk volt minden egyes terminusjelöltnél (1) annak hosszára, (2) azon elemek számára, amelyek tartalmazzák azt, és hosszabbak is nála, valamint (3) ez utóbbiak előfordulási számára. A C-érték kiszámítását ezért akkor végeztük el, amikor a mintaillesztési folyamat már lezárult, azaz az összes terminusjelölt a rendelkezésünkre állt.

A szövegkörnyezetet figyelembe vevő értékek közül a szintén a Frantzi és Ananiadou (1997) által kidolgozott súlyértéket vettük alapul, amihez megadtunk körülbelül 100 terminust. Kigyűjtöttük a környezetükben lévő tokeneket és megvizsgáltuk, hogy azok milyen gyakran fordulnak elő terminusok és nem terminusok közelében. A 100 terminus szövegkörnyezete által kiszámított értékek alapján történt a súlyérték számítása minden egyes terminusjelöltre az alábbi módon:

$$Weight(w) = 0{,}5 \cdot \left(\frac{t(w)}{n} + \frac{ft(w)}{f(w)} \right)$$

A képletben a w a környezetben előforduló szó, n azon terminusok száma, amelyekre ez a vizsgálat kiterjed (az esetünkben 100), $t(w)$ azon esetek száma, ahol egy terminus legalább egyszer ezzel a w szóval áll együtt. Az $ft(w)$ azt mutatja meg, hogy a w szó összesen hányszor fordul elő terminusokkal együtt, $f(w)$ pedig w korpuszbeli előfordulásainak száma.

A súly-, a C- és *weirdness*-értékek kiszámítása után azokból egy összevont mértéket hoztunk létre, amelyre igaz az, hogy amennyiben egy adott terminusjelölt ennek a mértéknek egy adott értékét átlépi, akkor az nagy valószínűséggel terminus. Ehhez (1) először a három mértéket egy adott tartományra kellett képezni, hogy egymással kompatibilisek legyenek. Ezt követően (2) meg kellett határozni, hogy az egyes mértékek mennyire voltak fajsúlyosak a

terminuskivonatolás során (pl. a *weirdness*-érték 60%-ban, a súly- és C-érték 20%-ban számítson bele az összevont értékbe). A három mérték egységes tartományba való leképezéséhez a 0 és 1 közötti irracionális tartományt választottuk, mert így a valószínűségszámításban használatos függvényekkel is számolhattunk, ugyanis a valószínűségi értékek mindig a [0;1] tartományban találhatók, ahol 0 a lehetetlen esemény valószínűségét mutatja, az 1 pedig a biztos eseményét.

Eredmények

Cabré és mtsai (2001) alapján elsőként azt feltételeztük, hogy a szabály alapú módszerekkel nagymértékű fedés érhető el alacsonyabb pontossággal. Az elemzett szabadalmi leírások alapján az eredményeink igazolták ezt: tisztán szabály alapú mintákkal az informatikai korpuszon 0,8-es fedést és a 0,53-os pontosságot (F-érték: 0,64) értünk el, a másik korpuszon ezen értékek hasonlóak, rendre 0,74; 0,54 és 0,62. A szabály alapú szűréssel mindkét korpuszon a fedés kevésbé, a pontosság értékei viszont jelentősen nőttek (kb. 0,09-dal).

A statisztikai módszerek alkalmazása előtt az első öt, informatikai szövegen végeztünk kísérletet. A három statisztikai mértéket szinte az összes lehetséges kombinációban kipróbáltuk, és megmértük, hogy az egyes kombinációkban mennyi lett az öt dokumentumban a terminusok ezen összevont értéke, majd megmértük, hogy mely esetekben lett maximális az összes dokumentum F-értéke. Ehhez külön kellett venni az egy- és a többszavas jelölteket, amelyek közül az utóbbiakra a C-érték is alkalmazható, míg az előbbire csak a *weirdness*- és a súlyérték.

Az egyszavas terminusok esetében a *weirdness*-érték a legmeghatározóbb: ha ezek fajsúlyosságát 0,8 vagy 0,9-re választjuk, akkor érhető el a legnagyobb F-érték. A többszavas terminusok esetén viszont éppen a súly értékei meghatározóbbak: minél nagyobbnak választjuk ezt meg, az F-érték annál inkább nő. A C-értékkel kapcsolatban meglepőek az eredmények: legtöbbször nem járult hozzá az F-érték növeléséhez.

Második hipotézisünket, miszerint a statisztikai módszerek a pontosságot nagymértékben növelik, ám a fedést csökkentik, szintén igazolták az adatok: ha az összevont érték (ÖÉ) küszöbét magasnak választottuk, azaz csak azon terminusjelölteket fogadtuk el, amelyeknél nagy volt az ÖÉ, akkor a pontosság jelentősen megnőtt, de a fedés (és ezáltal az F-érték is) természetesen jelentősen csökkent. Az informatikai korpuszban, ha a küszöböt 0,86-re választottuk, akkor a pontosság átlagban 0,9, az F-érték viszont ekkor csak 0,04. A legjobb F-értékeket akkor kaptuk, ha a küszöböt alacsonynak választottuk (0,03 és 0,05), ekkor az F-érték csak kis mértékben, kb. 1,5%-kal nőtt mindkét korpuszon a szabály alapú szűrést követően.

A saját terminológiakivonatoló eredményeit összevetettük más, szabadon felhasználható alkalmazásokkal. Az egyik ilyen a Jacquemin (2001) által kifejlesztett Fastr, a másik az Aubin és Hamon (2006) által létrehozott YaTeA (Yet Another Term ExtrActor).

1. táblázat: A Fastr, a YaTeA és a saját terminológiakivonatoló eredményeinek összehasonlítása a szabadalmi korpuszon

Alkalmazás	A23L			G06F		
	Fedés	Pontosság	F-érték	Fedés	Pontosság	F-érték
YaTeA	0,5711	0,3451	0,4270	0,5826	0,3045	0,3983
FastR	0,5764	0,4130	0,4806	0,5349	0,3962	0,4523
Saját	**0,7614**	**0,6176**	**0,6809**	**0,8280**	**0,6367**	**0,7185**

Az 1. táblázat eredményeiből jól látható, hogy mindkét korpuszon a saját alkalmazás eredményei mind fedés, mind pontosság tekintetében meghaladták a másik kettő alkalmazás eredményeit. Bár mindhárom alkalmazás szabály alapon működik, az eredmények azért is lehetnek

különbözőek, mert nem ugyanazt a morfoszintaktikai elemzőt használják, hanem a *TreeTaggert*[4]. A különbség másik oka lehet még, hogy ezen terminológiakivonatolókat általánosabb korpuszokra tervezték.

Hibaforrások, eredmények elemzése

A fedés, pontosság és F-értékek felsorolása nem lehet elegendő: célszerű ugyanis megvizsgálni, hogy az egyes módszereknél melyek voltak azok a terminusok, amelyek a szűrés ellenére mégsem kerültek be a terminusjelölt-listába, és melyek azok a nem terminusok, amelyek a különböző módszerek ellenére mégis bennmaradtak a listában. A főbb hibaforrások a következők: (1) a POS-tagger hibázott, (2) a szűrés ellenére sok nemterminusi főnévi elem maradt a terminusjelölt listában.

Ehhez mindkét szaknyelv (informatikai és alapvető emberi szükségletek) korpuszából választottunk egy-egy szabadalmi leírást, amelyben az összes hibás esetet végignéztük. A G06F korpuszból az FR2008051823[5] számú dokumentumot, az A23L korpuszból az FR2007051158[6] számú dokumentumot választottuk. Az első dokumentum 2760 szövegtokenből áll, és 193 kézzel annotált terminussal rendelkezik. A második dokumentum mérete 4372 szövegtoken, és a kézi annotációk alapján 462 különböző terminussal bír.

Mivel a terminológiakivonatoló elsősorban szabály alapú módszerekkel működik, azaz adott morfoszintaktikai mintákkal, ezért fontos, hogy az automatikusan működő POS-tagger minél hatékonyabban működjön. Azonban minden annotáló program egy kisebb-nagyobb hibaszázalékkal dolgozik, gyakori eset például a főnevek mellékneveknek jelölése, például a *solvant* 'oldat' helyett legtöbbször 'oldó' melléknévként lett jelölve. Az informatikai korpuszban a fel nem ismert terminusok közül (19) 5 esetben volt ez a hiba oka, a terminusjelölt-listába 6 került feleslegesen emiatt (76-ból). Az A23L korpuszon az arány másképp alakult, mert itt 31 esetben azért nem ismerte fel a terminust, mert rossz volt a *POS-tag*e (a fel nem ismert terminusok száma 99), és 17-szer került feleslegesen a listába (164 eset közül).

Természetesen nem minden terminus, ami a megadott szintaktikai mintára illeszkedik: például az egyszavas főnevek gyakran lehetnek köznyelvi egységek is. Az informatikai példaszöveg esetében 39 olyan terminusjelölt szerepel, ami valójában köznévi egység, az A23L szövegében pedig 52. Ilyenek a *place* 'hely', *an* 'év' stb.: ezen szavak aránya kb. 30%-os a hibásan terminusnak jelölt elemek közül.

A statisztikai módszerek az F-értéket csak kisebb mértékben tudták növelni. A G06F (informatikai) szövegben, ha az összevont értéket elég alacsonynak (0,05), választottuk, akkor a fedés értékének megtartásával a pontosságot sikerült a szabály alapú módszerhez képest 4%-kal növelni (0,5315-ről 0,5738-re). Ez azt jelenti, hogy az alacsony összevont értékkel rendelkező elemek eltávolításával a pontosság nőtt, tehát valóban nem terminusi elemeket távolítottunk el. Ezen nem terminusi elemek száma 15, és többek között olyan elemeket értünk rajta, mint *instant* 'pillanat', *lieu* és *place* 'hely', *arrêt* 'megállás'.

---

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[5] Canu, S., Grilheres, B., Brunessaux, S. (2009) *Méthode et système d'annotation de documents multimédia* (Multimédia dokumentumokat annotáló eszköz és rendszer, saját ford.)
http://www.wipo.int/patentscope/search/en/detail.jsf?docId=WO2009053613&recNum=1&maxRec=4&office=&prevFilter=&sortOption=Pub+Date+Desc&queryString=FP%3A%28Method+and+system+for+annotating+multimedia+documents%29&tab=PCTDescription

[6] Bourges, C. (2009) *Utilisation du safran et/ou du safranal et/ou de la crocine et/ou de la picrocrocine et/ou de leurs dérivés en tant qu'agent de satiété pour le traitement de la surcharge pondérale* (Sáfrány és/vagy safranal és/vagy krocin és/vagy pikrokrocin és/vagy származékaik használata jóllakottság érzését keltő szerként a túlsúly kezelésére, saját ford.)
http://www.wipo.int/patentscope/search/fr/detail.jsf?docId=EP15054358&recNum=1&office=&queryString=safran+safranal+crocine+picrocrocine+agent+de+sati%C3%A9t%C3%A9&prevFilter=&sortOption=Date+de+pub.+antichronologique&maxRec=2

További kutatási célok

A TE alkalmazás a későbbiekben kiterjeszthető más területen írt szabadalmi leírások esetére is, amelyekben akár más szerkezetű minták is előfordulhatnak, ezért az ezekre a korpuszokra kidolgozott minták bővítésre szorulhatnak. Érdemes még megnézni, hogy a terminusok statisztikai szűrése történhetne-e más hasonló mértékekkel is, amelyek részletes listája az 5. fejezetben olvasható. Ebben az esetben lehet, hogy a C-érték helyett más *unithood*-mértékkel jobb eredményt érhetünk el, és ezáltal ki tudjuk szűrni azokat a többszavas terminusokat, amelyekben például a terminus részét nem képző melléknévi utómódosító található. A 10-10 szabadalmi leírásból érdemes lenne egy olyan korpuszt összeállítani, ahol a terminusok nem külön listában szerepelnek, hanem a szövegkörnyezettel együtt, az eredeti szövegben bejelölve. Ez azért szükséges, mert egy ilyen korpusz már tanuló algoritmusoknak is alapul szolgálhat, és jelenleg nem érhető el terminológiai korpusz, csak terminológiai adatbázis, ahol a terminusok csak fel vannak sorolva. Érdemes lenne még megvizsgálni azt is, hogy más típusú korpuszokon milyen eredményt produkálnának ezek a módszerek: egy kevésbé szakmai szövegen, például didaktikusabb, magyarázó szövegek esetében valószínűleg a statisztikai módszerekre jobban szükség van, mert azon szövegekben több nem terminusi elem is előfordulhat.

Irodalom

Ahmad, K., Gillam, L., Tostevin, L. (1999). University of Surrey Participation in Trec8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (Wilder). In *The Eighth Text REtrieval Conference (TREC-8)*.

Ahmad, K. (2001). The role of specialist terminology in artificial intelligence and knowledge acquisition. In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*, 809–844.

Aubin, S., Hamon, Th. (2006). Improving term extraction with terminological resources. *Advances in Natural Language Processing Lecture Notes in Computer Science*, 380–387.

Bosredon B., Tamba I. (1991). *Verre à pied, moule à gaufres* : préposition et noms composés de sous-classe, *Langue française* **91**: 40–55.

Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Doktori disszertáció, Informatique Appliquées aux Sciences Humaines de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Cabré, M. T. (1999). *Terminology. Theory, methods and applications*, Amsterdam/ Philadelphia, John Benjamins.

Cabré, M. T., Bagot, R.E., Vivaldi Palatresi, J. (2001). Automatic term detection. A review of current systems. In Bourrigault, D., Jacquemin, Ch., L'Homme M-C. (eds.) *Recent advantages in Computational Terminology*. Amsterdam/Philadelphia, John Benjamins Publishing Co., 53–87.

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In Cinque, G. (ed.) *Path Towards Universal Grammar. Studies in Honor of Richard Kayne,* Washington: Georgetown University Press, 85–110.

Connexor, http://www.connexor.eu/technology/machinese/

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD-értkezés. Université de Paris VII, Paris.

Enguehard, Ch. (2005). Un banc de test pour la reconnaissance de termes en corpus. In Williams, G.. (ed.) *La linguistique de corpus*. Rennes, Presses Universiataires de Rennes, 273–286.

Frantzi, K.T., Ananiadou, S. (1997). Automatic term recognition using contextual clues. In *European Research Consortium for Informatics and Mathematics - Cross-Language Information Retrieval,* ERCIM-97-W003, 25–32.

Ha, L.A., Fernandez, G., Mitkov, R., Corpas, G., (2008). Mutual bilingual terminology extraction. In *Proceedings of LREC 2008* (CD-ROM), Marrakech, 1818–1824.

Jacquemin, Ch. (2001). *Spotting and discovering terms through natural language processing*, Cambridge, Cambridge(MA)/London, MIT Press.

Jacquemin, Ch., Bourrigault, D. (2003). Term extraction and automatic indexing. In Mitkov, R. (ed.) *The Oxford handbook of computational linguistics*, Oxford, Oxford university Press, 599–615.

Justeson, J. S., Katz, S. M., (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1): 9–27.

L'Homme, M-C. (2004). *La terminologie: principes et techniques*, Montréal, Les Presses de l'Université de Montréal.

Maynard, D., Ananiadou, S. (2000). Identifying Terms by their Family and Friends. In *Proceedings of COLING 2000*, Luxembourg, 530–536.

Nagy, Á. (2008). L'extraction terminologique : l'un des défis actuels de la linguistique informatique. In Kovács, K. és Nagy, Á. (szerk.) *Le Passé dans le Présent, le Présent dans le Passé.* Séminaire doctoral, Szeged, 2007. október 25–26., Szeged, JATEPress. 283–290.

Nagy, Á. (2009a). Kísérlet szintaktikai és statisztikai módszerekkel történő automatikus terminológiakivonatolásra francia nyelvű szövegekből. In Sinkovics, B. (szerk.) *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. JATEPress, 2009. 71–86.

Nagy, Á. (2009b). La structure interne des termes techniques du français et leur reconnaissance par ordinateur. In Kieliszczyk, A., Pilecka, E. (eds.), *La perspective interdisciplinaire des études françaises et francophones*. Łask, Oficyna Wydawnicza LEKSEM, 117–123.

Nagy, Á. (2012a). Contrasting French nominal terms to common language NPs – towards a rule-based term extractor. In Surányi, B. és Varga, D. (szerk.) *Proceedings of the First Central-European Conference for Postgraduate Students*, Piliscsaba, Pázmány Péter Katolikus Egyetem, 191–210.

Nagy, Á. (2012b). La comparaison de la structure interne des groupes nominaux ordinaires et des termes nominaux: les groupes prépositionnels et les groupes adjectivaux. In Gécseg, Zs., Penke, O., Szász, G. (szerk.) *Acta Romanica* (XXVIII). Szeged, JatePress, 7–18.

Petit, G. (2001). L'introuvable identité du terme technique, *Revue Française de Linguistique Appliquée* **VI**(2): 63–79.

Riegel, M., Pellat, J-Ch., Rioul, R. (2009). *Grammaire méthodique du français* (4. kiadás). Paris, PUF.

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam & Philadelphia, John Benjamins.

Sauron, V. A. (2002). Tearing out the terms: evaluating terms extractors. In *Proceedings of Translating and the Computer 24*, London, Britain.

Vasconcellos, M. (2001). Terminology and Machine translation, In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*, 697–723.

Wong, W., Liu, W., Bennamoun, M., (2008). Determination of unithood and termhood for term extraction. In Song, M. és Wu, Y. (eds.) *Handbook of Research on Text and Web Mining Technologies*, IGI Global.

Wüster, E. (1976). La théorie générale de la terminologie, un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et la science des objets. In *Actes du colloque international de terminologie*, Québec 5–8 octobre 1975, Québec, L'Éditeur officielle du Québec.

Wüster, E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In Rondeau, G. és Felber, H. (eds.) *Textes choisis de terminologie. Vol. I: Fondements théoriques de la terminologie*, Québec, Université Laval – GIRSTERM, 55-114.

A disszertáció témaköréhez kapcsolódó publikációk

Nagy, Á. (2012). Contrasting French nominal terms to common language NPs – towards a rule-based term extractor. In Surányi, B. és Varga, D. (szerk.) *Proceedings of the First Central-European Conference for Postgraduate Students*, Piliscsaba, Pázmány Péter Katolikus Egyetem, 191–210.

Nagy, Á. (2012). La comparaison de la structure interne des groupes nominaux ordinaires et des termes nominaux: les groupes prépositionnels et les groupes adjectivaux [A hagyományos főnévi csoportok és a főnévi terminusok belső szerkezetének összehasonlítása: A prepozíciós szintagmák és a melléknévi szintagmák]. In Gécseg, Zs., Penke, O., Szász, G. (szerk.) *Acta Romanica* (XXVIII). Szeged, JatePress, 7–18.

Nagy, Á. (2011). Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból. In Váradi, T. (szerk.) *V. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Budapest, MTA Nyelvtudományi Intézet, 107–114.

Nagy, Á. (2011). Vers une définition opératoire du *terme technique* [A *terminus technicus* műveleti definíciója felé]. In Oszetzky, É. és Krisztián, B. (eds.) *Cahiers francophones d'Europe Centre-Orientale 14. (Mots, discours, textes – Approches diverses de l'interculturalité francophone en Europe Centre-Orientale*, Pécs, Pécsi Tudományegyetem, 213–220.

Nagy, Á. (2010). L'extraction terminologique automatique: quel domaine de terminologie [Automatikus terminológiakivonatolás: melyik terminológiai terület]? In Kovács K. és Nagy Á (szerk.) *Acta Romanica. Tomus XXVII. Studia Iuvenum*, Szeged, JATEPress, 7–15.

Nagy, Á. (2010). Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból különböző módszerek segítségével. In Tanács A. és Vincze V. (szerk.), *MSZNY 2010*, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 375–378.

Nagy, Á. (2010). Reconnaissance de groupes nominaux: entre linguistique et intelligence artificielle [Főnévi csoportok felismerése: a nyelvészet és a mesterséges intelligencia határán]. In Malinovská Z. (szerk.), *Parenté/s*. Prešov, Filozofická fakulta Prešovskej univerzity v Prešove, 215–221.

Nagy, Á. (2009). Kísérlet szintaktikai és statisztikai módszerekkel történő automatikus terminológiakivonatolásra francia nyelvű szövegekből. In Sinkovics B. (szerk.), *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. JATEPress, 71–86.

Nagy, Á. (2009). La structure interne des termes techniques du français et leur reconnaissance par ordinateur [A terminusok belső szerkezete és számítógéppel történő felismerése]. In: Kieliszczyk A. és Pilecka E. (eds.), *La perspective interdisciplinaire des études françaises et francophones*, Łask, Oficyna Wydawnicza LEKSEM, 117–123.

Nagy, Á. (2009). Les premiers pas vers la création d'un extracteur automatique de groupes nominaux à partir de textes hongrois étiquetés [Egy annotált magyar szövegeken működő automatikus főnévicsoport-kinyerő rendszer létrehozásának első lépései], In *Acta Romanica, Tomus XXVI*. Szeged, JATEPress, 41–48.

Nagy, Á. (2008). L'extraction terminologique: l'un des défis actuels de la linguistique informatique [Terminológiakivonatolás: a számítógépes nyelvészet egyik aktuális kihívása]. In Kovács K. és Nagy Á (szerk.), *Le Passé dans le Présent, le Présent dans le Passé. Séminaire doctoral*, Szeged, JATEPress, 283–290.