KÍSÉRLET SZINTAKTIKAI ÉS STATISZTIKAI MÓDSZEREKKEL TÖRTÉNŐ AUTOMATIKUS TERMINOLÓGIAKIVONATOLÁSRA FRANCIA NYELVŰ SZÖVEGEKBŐL

NAGY ÁGOSTON

A terminológiakivonatoló alkalmazások leginkább csak más programok kiegészítéseként használatosak (mint ahogy a helyesírás-ellenőrzők is), tehát önálló alkalmazásként csak ritkábban, ennek ellenére ez a számítógépes nyelvészet egyik igen kutatott területe. A terminológiakivonatolás egyik alkalmazási célja az automatikus szövegindexelés, amely során egy megadott szöveges fájl rá jellemző fontosabb kifejezéseit kivonatoljuk. Erre az eszközre épülnek például az internetes keresőmotorok egy része is, amelyek a már indexelt honlapokat tárolják a gyorsabb és hatékonyabb keresés megvalósítása érdekében (Enguehard 2005). Ugyanakkor terminológiakivonatoló eszközökre épülhetnek gépi fordítást megvalósító alkalmazások (pl. Vasconcellos 2001), illetve információkinyerő eszközök is (l. Ahmad 2001).

A terminológiakivonatolás egy másik fontosabb felhasználási területe a fordítói munka elősegítése. Ha például egy többszáz oldalas szakmai könyvet (mondjuk egy szoftver vagy hardver leírását) kell rövid idő alatt konzekvensen lefordítani, azt általában egy fordító nem tudja megoldani. Ilyenkor fontos lehet egy terminusjelölt lista, amelyet ha a lektor előre kézhez kap, akkor már a szöveg fordítóknak történő kiadása előtt lehetősége lenne megadni az abban szereplő terminusok idegen megfelelőjét. Így a csoportban dolgozó fordítók biztos minden terminust ugyanúgy fognak fordítani.

A jelen cikknek több célja van: az egyik a terminológiakivonatolás által felvetett problémák és egyben annak a bemutatása, hogy milyen nehézségekbe ütközhetünk egy terminológiakivonatoló eszköz megalkotásakor, valamint egy saját fejlesztésű terminológia-kivonatoló eszköz működési elvének leírása. Jóllehet sokféle szófajú terminus létezik, ez az alkalmazás kizárólag a főnévi terminusok kinyerésére összpontosít, mint ahogy a legtöbb terminológiakivonatoló eszköz. A cikk első részében a terminus fogalmának definiálási nehézségeiről esik szó, hiszen ez a fogalom is csak első látásra tűnik egyszerűnek, a valóságban azonban itt is nehézségekbe ütközünk. Ez-

után különböző terminológiakivonatoló eszközök rövidebb bemutatása következik; ez azért lényeges mert az általunk kidolgozott terminológiakivonatoló alkalmazást is ezek alapján hoztuk létre. A saját terminológiakivonatoló alkalmazás részletes bemutatása előtt a tesztkorpusz rövid leírása található. A cikk legvégén pedig kiértékeljük az alkalmazást az erények és hibák függvényében.

A terminus fogalma

Egy terminológiakivonatoló alkalmazás megvalósításakor felmerül. hogy először a terminus fogalmát kell a lehető legpontosabban meghatározni. Azonban a terminológiával kapcsolatos szakirodalom alapján azt állíthatjuk, hogy ez a feladat korántsem olyan egyszerű, mint ahogy az első látásra tűnik. Azaz rengeteg definíció létezik, amelyek pontossága és helytállósága a kérdéses publikáció céljától is függ: nem mindegy ugyanis, hogy az adott cikk ezt a kérdést mint terminológiai (elméleti vagy gyakorlati) vagy nyelvészeti szemszögből közelíti meg. E két nézőponti különbséget jól szemlélteti például Petit (2001).

A jelen cikk a terminusok definíciójának terminológiai nézőpontját tükrözi, mert célunkhoz, tehát a terminológiakivonatoláshoz, ez áll legközelebb. A klasszikus (tehát szemantikai alapú) nézőpont szerint a terminust az alábbi kritériumok alapján írhatjuk le: a terminushoz mindig tartozik egy fogalom (*concept*), amelyet a terminus egyértelműen elnevez, és a terminus mindig köthető valamilyen szakterülethez (Cabré 1999; Wüster 1976; Sager 2000). Ez matematikai fogalmakkal azt jelenti, hogy terminusok és fogalmak halmaza között bijektív leképezés van, tehát egy terminushoz (elvileg) egy és csakis egy fogalom tartozik, és egy fogalomhoz pedig egy és csakis egy terminus. Természetesen ez a kapcsolat csak konkrét szakterületen belül működik, hiszen például a *page* szó a szoftverek világában is több dolgot jelöl: míg a Wordben inkább *oldal,* addig a honlapok esetében *lap*.

Egy szövegegység akkor válik terminológiává, amint egy szakterülethez tartozó szemantikai hálózatba kerül. Itt megjegyzendő azonban, hogy ennél a pontnál is elég sok probléma merül fel, hiszen az, hogy mi tartozik egy szakterülethez és mi nem, azt néha nagyon nehéz eldönteni. Ennek oka, hogy a köznyelv és a szaknyelv (ez utóbbiak közötti különbségről a következő részben foglalkozunk bővebben) állandó kapcsolatban áll egymással. A szaknyelv tehát a köznyelvből vesz át szavakat, de ez fordított irányban is működik; erre példa lehetne a *háló* kifejezés, amely az informatikai szak-

nyelvben (sőt ma már a köznyelvben is) az Internetet jelöli. A szaknyelv és a köznyelv állandó kapcsolatban van tehát, a kettőjük közötti kapcsolatot leginkább a metszet és egymásba ágyazás kifejezésekkel írhatjuk le (Petit 2001). Másik probléma a szakterületek közötti átfedés. Nagyon sok kifejezés átkerül az egyik szaknyelvből a másikba. Ez a mi esetünkben is probléma volt, hiszen az *analyse coûts-bénéfices* (költség-haszon elemzés) kifejezés a mi, informatikai témájú korpuszunkban is szerepelt, és terminus is, de mégsem ehhez a szakterülethez köthető.

A terminus ismertetőjegyei

Terminológiai szemszögből a terminus a köznyelvi szó ellentétje. A továbbiakban a terminus és a köznyelvi szó közötti különbségeket tekintjük át, ami azért fontos, mert ezek azok a kritériumok, amelyek alapján az ellenőrzési fázisban megállapíthatjuk, hogy az adott kivonatolt terminus-jelölt valóban az-e. A különbségek leírásához Sager (2000) és Cabré (1998) publikációit vettük alapul.

Legfontosabb különbség a terminus és a köznyelvi szó között, hogy a terminus egy szaknyelvhez köthető. A köznyelv és szaknyelv közötti különbség pedig abban áll, hogy az elsőt folyamatosan és nem tudatosan sajátítjuk el, míg a másodikat a köznyelv elsajátítása után tanuljuk meg tudatosan.

A köznyelvi szó létező és nem létező vagy képzeletbeli entitásokat is leírhat. Azonban a terminus leginkább létező egységeket ír le, és mindezt az adott szakterület képviselői közötti kommunikáció elősegítése végett.

Különbség adódhat még a két elem legitimizálásával kapcsolatban is. Egy köznyelvi szó legitimizálásához elég, ha azt az adott nyelvi közösség elfogadja, majd utána használja és megérti. Használata spontán, tehát külső intézmény azt nem befolyásolhatja (de mint tudjuk, gyakran ebbe is próbálnak beleavatkozni). A terminus akkor válik legitimmé, ha azt egy arra jogosult személy vagy intézmény hitelesíti, például egy tudományos intézet.

A terminus univerzális, tehát egy-egy fogalomhoz elvileg létezhetne minden nyelven terminus. A szó mindig egy konkrét nyelvhez tartozik, és két külön nyelvben nehéz találni olyan megfeleltetést egy-egy szó között, amelyek a két nyelvben minden kontextusban ugyanúgy használhatóak.

A gyakran többértelmű köznyelvi szavak jelentései a szövegkörnyezetből kövezethetőek ki, a kontextus ismerete nélkül nem mindig lehet egy adott szövegrészletnek jelentést adni. A terminus ezzel szemben univerzális,

és ezáltal jelentése magától értetődő és gyakran kontextus ismerete nélkül is megadható.

A szavak alapértelmezés szerint többértelműek, míg a terminusok nem. A szavak bizonyos stílushoz tartozhatnak, a választást mindig a nyelvi regiszter szintje adja meg. A terminusnak ezzel szemben nincsenek változatai.

A terminus definíciója – tisztán gyakorlati szemszögből

Kis (2005) szerint a terminológia által elfogadott hagyományos definíciók a terminusok pusztán formális tulajdonságait próbálják leírni a szövegbeli előfordulás figyelembevétele nélkül. A terminológia szerint például egy terminus akkor létezhet, ha az korábban explicit definiálásra és legitimizálásra került, és ezáltal egy adott szakterületen belül azt mindig konzekvensen használják. Azonban a valóságban, fordítói tapasztalat alapján, állítható, hogy terminusok egy szövegben előzetes definíció vagy legitimizálás nélkül is bármikor megjelenhetnek. Azonban, fordítási szempontból lényeges, hogy a terminusokat a fordításokat végző személy konzekvensen adja vissza, mind szemantikai, mind formai szemszögből.

A hagyományos terminológiai nézőpont azért sem válhat sikeressé a fordítók körében, mert őket nem érdekli a terminus eredete. Egyetlen céljuk van: az adott terminus helyes és konzekvens fordítása (ami valljuk be, több fordítót magába foglaló projektek esetében igencsak időigényes lehet a folytonos egyeztetések miatt). Kis (2005) szerint tehát a fordítás szemszögéből nézve a terminusnak egy fő tulajdonsága van: mindig következetesen kell fordítani.

Kis és mások (2004) szerint a terminus egy szövegnyelvészeti jelenség, amelyet két fő jellemző határoz meg:

(i) terminológiai helyzet: a terminológiai helyzet olyan űrként jelenik meg a szövegben, amelyet terminussal kell megtölteni; ebbe a helyzetbe az olvasó egy terminust vár.
(ii) terminológiai szerep: ha egy szó a fent említett helyzetben van, akkor terminusként értékelődik. Ebben a helyzetben az adott szó a köznyelvi helyzettől eltérően viselkedik, felveszi a terminusra vonatkozó tulajdonságokat.

A terminológiakivonatolók működési elvei

A terminológiakivonatolás célja olyan szavak és kifejezések listájának létrehozása, amelyek lehetséges terminusok. Azonban vizsgálatok kimutatták, hogy ezen alkalmazások kimenete nem tekinthető teljesen biztosnak emberi utóellenőrzés nélkül, ezért a kimeneti lista elemeire a „terminusjelölt" kifejezést alkalmazzuk a továbbiakban (Jacquemin 2001).

Alapjában véve két fő módszer létezik: a statisztikai alapú és a szabályalapú (tehát nyelvészeti alapú) terminológiakivonatolók. Ez azonban nem azt jelenti, hogy ezen alkalmazások csak az egyikre támaszkodnának: a többségük a kettőt ötvözi (Maynard & Ananiadou 2001). Cabré et *al.* (2001) szerint nem is javasolt, hogy a két módszer közül csak az egyikre támaszkodjunk, mivel a szabályalapú kivonatolók túl nagy zajt okoznak (tehát a kivonatolt terminusjelöltek száma magasabb, mint a valós terminusoké), a statisztikai alapúak pedig túl nagy csendet (a terminusjelöltek listája sok terminust nem tartalmaz).

A hagyományos megközelítés a szabály alapú megközelítést támogatja: a terminusok a belső morfoszintaktikai szerkezetük segítségével vonatolhatók ki, ugyanis vannak olyan szintaktikai minták, amelyekre csak a terminusok illeszkednek. Ez a francia nyelvre különösen jellemző, ezért a saját kivonatolónk is elsősorban erre épít. Ez azonban leginkább csak a gazdasági és informatikai szövegek esetén alkalmazható, egy filozófiai szakszövegben például már nehezebb dolgunk lenne, mivel az ilyen típusú szövegek sokkal inkább szabálykövetőek, így ezekben a nem-terminusok szerkezete megegyezik a terminusokéval. A gazdasági és informatikai nyelvre pedig az egyszerűsítés jellemző, amely miatt a terminusok szerkezete egyszerűbb lesz a köznyelvi főnévi csoportokhoz képest.[1] A nyelvészeti megközelítés tehát a szabályos kifejezések és a véges állapotú automaták technikájára épül (Cabré et *al.* 2001).

A statisztikai kivonatolók olyan lexikai egységeket keresnek fel a szövegekben, amelyek gyakrabban fordulnak elő együtt a többi egységhez képest. Ezt Kis (2005) még ki is egészíti azzal, hogy ezután minden nemüres szópárra meg kell mérni annak gyakoriságát egy hétköznapi nyelvet tükröző referenciakorpuszban. Ha a gyakoriság nagyon eltér, akkor biztos terminusról beszélhetünk. A statisztikai alkalmazások gyakran használnak még asszociá-

[1] Leggyakoribb ilyen eset a prepozíció elhagyása, például *doseuse produits aigres* (savanyított termék töltőgép) a szabályt követő *doseuse **de** produits aigres* helyett.

ciós és távolsági mértéket is. Ugyanis azok az igazi terminusok, amelyek egy szövegben egymáshoz mindig közelebb vannak, illetve gyakrabban fordulnak elő együtt, mint külön.[2]

A terminológiakivonatolók sokasága mind azt mutatja, hogy a terminológiakivonatolás korántsem egyszerű feladat, és a szakirodalmat áttekintve az is kiderülhet, hogy minden kivonatoló vagy egy-egy speciális szakterületre vagy nyelvre van kidolgozva. Különbség adódik még az „extrák" tekintetében is. Többek között ezek alapján fogjuk bemutatni a különböző kivonatolókat.

Az egyik leghíresebb és legtöbbet idézett kivonatoló a FASTR (Jacquemin 2001). Ez az alkalmazás nemcsak a terminusok kivonatolására képes, de nagy előnye, hogy azok variánsait is képes felfedezni, így hatékonyabb is. Jacquemin (2001) ehhez a szintaktikából már jól ismert újraírószabályokat használ, amelyeket ő metaszabályoknak nevez. Ezek különböző nyelvtani jelenségeket fednek le, léteznek ugyanis mellérendelő, beszúró és permutációs szabályok. A mellérendelés egyik szabálya a:

$$\text{Metarule Coor}(X_1 \rightarrow X_2\ X_3) = X_2\ C_4\ X_5\ X_3,$$

ahol C_4 egy mellérendelő kötőszó (*conjunction*), X pedig egy nyelvtani kategória (nagy valószínűséggel N). Ezen szabály segítségével például megtalálhatjuk a *serum albumin* (savófehérje) variánsait, mint például *egg and serum albumin* (tojás- és savófehérje).

A LEXTER (Bourrigault 1994) is említést érdemel: elsősorban a franciaországi áramszolgáltató vállalat, az EDF dokumentumainak indexelése végett hozták létre. Ennek a kivonatolónak nagy előnye, hogy képes helyesen elhatárolni a főnévi csoportokat: először maximális hosszúsági főnévi csoportokat kivonatol, amelyeket aztán később feldarabol kisebb egységekre, ha szükséges. Ezt a vágást a prepozíció bal oldali kontextusa dönti el, amely a főnévi csoport lehetséges határa.

Kis (2005) alkalmazása a kontextust is figyelembe veszi. A szerző ugyanis észrevette, hogy egy lapos mondatelemző segítségével újabb terminusokat lehet felfedezni. A magyarban például a fókuszban lévő kifejezések nagyobb valószínűséggel válnak terminussá. Másik előnye ennek a kivonatolónak egy köznyelvi használatot tükröző referenciakorpusz használata. Ha a

[2] Az asszociációs mértékekről bővebben Daille (1994)-ben.

terminusjelölt ebben nem vagy csak alig szerepel, akkor sokkal valószínűbb, hogy az terminus.

Végül érdemes még megemlítenünk Meilland & Bellot (2005) kivonatolóját is, amelyet rövid termékismertetők terminusainak kivonatolására alkottak. Mint ahogy azt mindenki beláthatja, ezek a termékismertetők rövidek, tömörek, szóismétléstől mentesek, tehát a nagy korpuszokon hatékony alkalmazások itt nem sokat érnek. Ők ki is mutatták, hogy ezen szövegtípus esetén a terminusok gyakoriságára építő alkalmazások vajmi keveset érnek, de az asszociációs mértékek használata nagy fedést biztosít. Ezen asszociációs mértékek kiszámítása matematikai alapon történik, a leggyakoribb mérték például a tapasztalati korrigált szórásnégyzet képletét alkalmazza, amelyet két adott szövegelem távolságainak értékéből számítunk.

Tesztkorpusz

Tesztkorpuszként egy nagyon széles körben ismert informatikai e-könyvet választottunk, amely az objektumorientált programozásba vezeti be az érdeklődőket. Ez egészen pontosan Bruce Eckel *Thinking in Java* című műve, illetve annak is francia nyelvű változata, *Penser en java*. Jogos a kérdés, hogy miért választunk angolról fordított, és interneten elérhető művet korpuszként, de a választást számos érv támasztja alá. Például számos egyetemi kurzus esetén van feltüntetve hivatalos és megbízható szakirodalomként, valamint ez az a mű, amely alapján sokan elkezdik a Java nyelv tanulását, illetve amely alapján sokan elmélyítik programozói ismeretüket. A második ok, hogy ez az a terület, amely esetében mi magunk is könnyebben tudjuk eldönteni, hogy mi terminus és mi nem, ez pedig az ellenőrzési szakaszban különösen fontos.

A korpuszt ezen belül is a könyv első fejezete alkotja, mégpedig az *Introduction sur les objets* [Bevezetés az objektumokba] című rész. Azért ez a fejezet, mert itt szinte az összes terminus megjelenik, amely az objektumorientált programozáshoz köthető.

A korpusz először az Unitex v1.2. programmal és annak kézzel megírt egyértelműsítő gráfjaival lett szótövesítve, majd a szavak szófaj szerint meg lettek címkézve. Fontos megjegyezni, hogy a könnyebb kezelhetőség érdekében a Unicode (2B/karakter) kódolás helyett a szöveget ASCII formátumra (1B/karakter) kellett konvertálni, mert a használt programozási nyelv (Perl), az előbbit nem támogatja. Így ugyan az ékezetes karakterek egy része elveszett, de ez nem okozott kivonatolási hibát.

A kivonatolás módszere

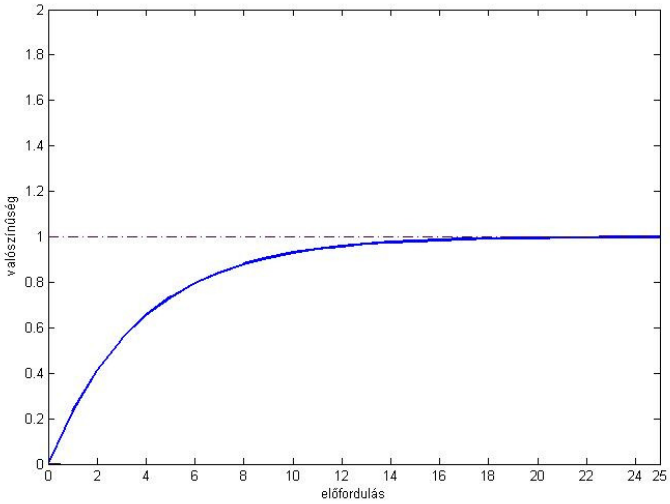
A terminológiakivonatoláshoz mi is követjük azt a megoldást, amit a legtöbb ilyen jellegű alkalmazás követ, tehát az alkalmazásunk kombinálja a szabályalapú és a statisztikai módszerek előnyeit, de elsősorban szabályalapú. A terminuskivonatoló alkalmazás Perl nyelven íródott, mivel ez az egyik legelterjedtebb és legkönnyebben kezelhető alkalmazás, amellyel mintát lehet illeszteni. A szintaktikai mintákat a programnak előre megadjuk, és az annotált korpuszból ő automatikusan kiválogatja a terminusjelölteket. Ezután a jelöltek egy valószínűségi értéket kapnak, amely több tényezőből áll össze. Az első, és egyben legegyértelműbb ilyen tényező, a belső morfoszintaktikai szerkezet, ugyanis észrevettük, hogy bizonyos szerkezetű főnévi frázisok nagyobb valószínűséggel válnak terminussá, mint mások. A franciában például majdnem biztosan terminus a N-N-N mintára illeszkedő csoportok: ezek biztosan terminusnak tekinthetők. Az olyan szerkezeteknél pedig, amelyek nem tipikus terminusszerkezetek, akkor ott a statisztikai módszerekhez folyamodtunk, tehát a program az előfordulási gyakoriságuk alapján döntött azok terminus jellegéről. Hiszen azt vehetjük észre, hogy minél gyakrabban fordul elő egy kifejezés egy adott szövegen belül, annál biztosabb, hogy az terminus. A *serveur* (szerver) szó például 52 előfordulással büszkélkedhet, míg a *vélo* (bicikli) csak eggyel. A gyakoriság minél inkább nő, annál nagyobb a valószínűsége, hogy az terminus, ezért a valószínűségi értéket a legjobban az exponenciális eloszlás függvénye írja le. Azonban ezt az eloszlásfüggvényt annyival módosítottuk, hogy az eredeti $P(\xi \leq x)$ helyére $P(\xi = x)$-et írtunk, mivel jelen esetünkben nem folytonos, hanem diszkrét valószínűségi változóról van szó, hiszen minden egyes előfordulási gyakoriságnál arra vagyunk kíváncsiak, hogy ahhoz a konkrét előforduláshoz milyen valószínűségi érték tartozik. A képlet tehát:

$$F(x) = P(\xi = x) = 1 - e^{-\lambda x}$$

(ha $x \to \infty$, akkor $P(x) \to 1$),
ahol x az adott főnévi csoport gyakorisága, λ az eloszlás paramétere, értéke pedig $\lambda = \frac{1}{E(x)}$. Fontos, hogy ezt az eloszlásfüggvényt kategóriánként kell kiszámítani, hiszen más gyakorisággal szerepelnek terminusok az N-A és más az N-P-N főnévi csoportok esetében. Mivel ez a paraméter ismeretlen,

ezért azt maximum likelihood módszerrel számíthatjuk ki. Ehhez vesszük a korpusz adott kategóriába beillő terminusait, majd a maximum likelihood módszer alapján behelyettesítjük azok előfordulásait.

Az N-P-N kategória esetén számításunk alapján $\lambda = 0{,}141$. Az ehhez tartozó eloszlásfüggvényt az alábbi ábra jelképezi:



1. ábra
Az N PREP N kategóriához tartozó főnévi csoportok előfordulási számához tartozó valószínűség

Íme néhány példa a különböző morfoszintaktikai szerkezetekhez tartozó valószínűségi értékekről, illetve egy-egy konkrét példa minden kategóriához:

Minta	Terminus	Nem terminus	Valószínűség
N N N	*programmation côté serveur* (szerver oldali programozás)	-	100%
N PREP N A N	*langage de programmation orienté objet* (objektumorientált programozási nyelv)	-	100%
N PREP N A	*nombre en virgule flottante* (lebegőpontos szám) *hiérarchie de classes unique* (egyedi osztályhierarchia)	*partie de code utile* (hasznos programrészlet)	75%
N PREP N	*durée de vie* (élettartam) *chaîne de décision* (döntési lánc)	*point de vue* (nézőpont)	75%
N A N	*programmation orientée objet* (objektumorientált programozás)	-	100%
N PREP INF	*code à exécuter* (végrehajtandó kód)	*capacité à étendre* (kiterjeszthetőségi képesség)	10%
N N N	*Common Gateway Interface*[3]	*mot clef interface* (interfész kulcsszó)	86%
N A	*classe abstraite* (absztrakt osztály) *héritage multiple* (többszörös öröklődés)	*module requis* (megkívánt modul) *machine lente* (lassú gép)	38%
N N	*ramasse-miettes* (szemétgyűjtő) *navigateur web* (webböngésző)	*mot clef* (kulcsszó)	98%
N	*disque* (lemez) *objet* (objektum)	*arbre* (fa) *cas* (eset)	70%

1. táblázat
Példák kategóriánként terminusokra és nem-terminusokra

[3] Az idegen szavak az egyszerűség kedvéért mind főnévként lettek annotálva.

Ezen a ponton meg kell jegyeznünk, hogy alkalmaztuk még Kis (2005) ötletét, még ha egy kicsit más formában is. Ő referenciaszótár segítségével szűr ki olyan kifejezéseket, amelyek egy szövegben gyakran előfordulnak, de mégsem terminusok. Egy hasonló szótárt mi is használunk, de ennek tagjait mi állítottuk össze. Ha az ebben szereplő szavak valamelyike szerepel a kivonatolt terminusban, akkor azt a rész töröljük a jelöltből, majd a főnévi csoport maradék tagjaival folytatjuk a kivonatolást. Ilyen szavak vagy kifejezések például az *ensemble de …* (vminek az együttese), *centaine de …* (mintegy száz …), *gens* (emberek) stb.

Nagyon fontos szempont volt még a tipográfiai jelek figyelembevétele az eredeti szövegben, ugyanis az idézőjelben, dőlt vagy félkövér karakterrel szedett szavak valószínűleg terminusok (pl. *interface, abstract*).

Ezen kívül létrehoztunk még egy olyan listát, amely azokból a szavakból áll, amelyet az Unitex szoftver szótövesítője és POS-címkézője nem ismert fel, mert az nem szerepelt annak szótárában. Ezek általában mind angol eredetű lexémák voltak vagy mozaikszavak, mint például *plug-in, LinkedList, applet,* stb. Itt meg kell jegyeznünk, hogy ezeket automatikusan terminusként ismertük el, vagy ha ezek terminusjelöltben szerepeltek, akkor azokat a jelölteket is terminusnak vettük.

Ezeken kívül létrehoztunk még egy másik listát is, amely az összes olyan „gyanús" kifejezést tartalmazta, amelyek szinte soha nem részei terminusoknak, de ha egy főnévi csoportban szerepel, akkor az utánuk következő szó vagy kifejezés nagy valószínűséggel terminus. Gyanús kifejezés például a *mot clef* (kulcsszó), *notion de* (vminek a fogalma), stb.

A FASTR (Jacquemin 2001) mintájára mi is próbáltuk figyelembe venni a különböző terminusvariánsokat. A leggyakoribb eset például a melléknévi csoport határozó általi kibővülése volt. Hasonló esetekben, a határozószót töröltük, és a határozószó nélküli főnévi csoportot besoroltuk a neki megfelelő kategóriába. Ezért került be például az *approche purement orientée objet* (tisztán objektumorientált megközelítés) az N Adv A N nemlétező terminuskategória helyett az N A N kategóriába *approche orientée objet* 0(objektumorientált megközelítés) címszóként.

Összegzésképpen, először maximális hosszúságú, és a terminusok mintájára illeszkedő, főnévi csoportokat kivonatoltunk. Ezután eltávolítottuk a referenciaszótárban is szereplő elemeket, így eltávolítottuk azokat a szavakat, amelyek terminusban nem szerepelhetnek. Ezután, ha a főnévi csoport „gyanús" kifejezéseket tartalmazott, akkor azt róla levágtuk, és a megmaradt

rész szinte maximális valószínűségi értékkel bekerült a terminusjelöltek listájába. Ezután minden terminusjelölthöz valószínűségi értéket rendeltünk, attól függően, hogy milyen gyakran fordultak elő a szövegben. Majd végül a különböző valószínűségi értékekből egyet hoztunk létre, amelyet minden jelölthöz hozzárendeltünk.

A valószínűségi értékek alapján pedig minden kivonatolt kifejezést bevettünk a terminusjelöltek listájába, ha egy bizonyos intervallumon belül helyezkedett el. Az intervallum meghatározásához kiválasztottunk egy száz terminust tartalmazó reprezentatív mintát. Mivel a valószínűségi értékek szórása ebben a mintában ismeretlen, ezért ez egy n-1 szabadsági fokú Student-eloszlás, amelyhez könnyedén számolhatunk konfidenciaintervallumot a megfelelő képlettel:

$$[E_n(\xi) - x_\alpha \frac{\sigma}{\sqrt{n}}, E_n(\xi) + x_\alpha \frac{\sigma}{\sqrt{n}}] \cdot$$

Ha a kivonatolt elem valószínűségi értéke ebbe az intervallumba beleesik, akkor bekerül a terminusjelöltek halmazába. Mi egy 90%-os konfidenciaintervallum mellett döntöttünk a minta alapján.

Eredmények

A számítógépes nyelvészeti alkalmazások hatékonyságának elemzéséhez általában két fő koefficienst használunk, az egyik a fedés (*recall*), a másik a pontosság (*precision*). A terminológiakivonatolás esetében a pontosság az eredményesen kivonatolt terminusok számának és a kivonatolt elemek számának hányadosa. A fedés pedig az összes kivonatolt terminusok számának és a valós terminusok aránya (Kis 2005).

Célunk egy hozzávetőlegesen magas fedés elérése volt, tehát hogy az összes lehetséges terminust kivonatoljuk, még ha a terminusjelölt listába olyan elemek is kerülhetnek, amik nem terminusok. Ehhez nagyban hozzájárult, hogy az alkalmazásunk nagyobbrészt szabályalapú, így az a fedés értékét még jobban növelte. A fedés növelése végett bizonyos szintaktikai csoportoknál, ahol magas volt a terminusok előfordulásának valószínűsége, ott azok gyakorisági értékét nem vettük figyelembe, hogy ne veszítsünk el terminust, még ha ez a pontosság kárára is válik. Így sikerült is egy 85%-os fedést elérni, amely nagyban annak is betudható, hogy a francia informatikai és gazdasági szaknyelv terminusainak szerkezete igencsak különbözik a hétköznapi főnévi csoportokétól.

Korpusz mérete	15382 szó
ebből főnevek száma	3950
Valós terminusok száma	649
Terminus-jelöltek száma	784
Összes kivonatolt terminus száma	545
Észre nem vett terminusok száma	135
Nem terminusok száma a terminus-jelöltek között	239
Fedés	82,78%
Pontosság	69,51%

2. táblázat
Az alkalmazás pontos eredményei

Hibaforrások

Azt mondhatjuk, hogy a legjelentősebb hibaforrást a korpusz hozzávetőlegesen kis mérete okozta, hiszen a valószínűségi arányok tekintetében hátrányban lévő szintaktikai csoportok esetében (például szimpla N, vagy N-A) főleg a kivonatolt elemek gyakorisága döntött azok terminus voltáról, így a ritkábban előforduló szakkifejezések nem szerepeltek a kimeneti listában.

Hibaforrás volt még a determinánsok kizárása, ami mellett korábban azért döntöttünk, mert a francia nyelv szakkifejezései ritkán engedik meg belső determináns előfordulását (a terminus előtt persze lehet). Ha azokat megengedtük volna, a hosszú NP-k esetében a bennük előforduló terminusokat eltakarva kivonatolta volna azokat a program, így a valós terminusok előfordulását csökkentette volna, és ezáltal a fedést is. Azonban szerencsére csak két, determinánst tartalmazó főnévi terminusunk volt, a *traitement des exceptions* (kivételkezelés) és a *traitement des erreurs* (hibakezelés), ahol a *des* a *de* prepozíció és a *les* határozott névelő összevont alakja.

A harmadik problémát a minták jelentik, hiszen a statisztikai alapú rendszerekkel ellentétben ez az alkalmazás elsősorban a mintákra épít. Azonban szinte lehetetlennek tűnik az a feladat, hogy az összes létező mintát felsoroljuk. Ráadásul kevésbé valószínű, hogy az adott korpuszon kívül bárhol máshol megjelenik egy N-Pro-PREP-Pro, márpedig a szövegben volt ilyen: *relation un-à-un* (egy-az-egyhez kapcsolat).

További hibaforrás volt például az eleve a pontosság növelése végett használt gyanús kifejezések listája, amely például tartalmazta a *type de* (egy-

fajta …) kifejezést; az ezután következő elemet automatikusan terminusnak vettük. Azonban a szövegben (és egyben a Java nyelvben) bőven akad *type de* kezdetű terminus, amelynek az előbbi szerkezet is a része, például *type de donnée abstrait* (absztrakt adattípus). Így ezek lekerültek a terminusjelölt listáról, vagy máshogy kerültek oda be, mint például *donnée abstrait* (absztrakt adat).

Ezen kívül probléma volt az is, hogy a terminológiakivonatoló program pontossága, szabály alapú mivolta miatt, nem volt megfelelő mértékű a magas fedés mellett, de egy statisztikai alapú programmal pontosan az ellenkezőjét értük volna el, a mi célunk pedig elsősorban a terminusok többségének kinyerése volt.

Gond lehet még a terminus fogalmának értékelésével is: mint ahogy egy korábbi alfejezetben is elhangzott, nem mindig evidens, hogy mi terminus és mi nem. Ezt nem mindig volt könnyű eldönteni, és itt fontos megjegyezni, hogy mindig a definíciók alapján próbáltuk ezt megítélni, és szerencsére csak egy-két esetben kellett önkényes döntést hoznunk.

Az utolsó problematikát igazából az jelentené, ha ezt a programot más nyelvű korpuszokra alkalmaznánk, és ez egyben az összes szabályalapú kivonatoló alkalmazás problémája is, hiszen ugyanaz a szintaktikai minta valószínűleg más nyelveken nem köthető terminusokhoz, már ha azokban a nyelvekben egyáltalán előfordulhat az a minta.

Összegzés

Célunk egy olyan terminológiakivonatoló eszköz kidolgozása volt, amely francia nyelvű, szótövesített és annotált, informatikai szakszövegekből képes az abban található főnévi terminusok kinyerésére. Előtte azonban áttekintettük, hogy egyáltalán mik azok a formai definíciók, amelyek alapján a terminusokat fel lehet ismerni egy adott szövegben: ez különösen a végső fázisban volt fontos, ahol a programot kiértékeltük, hiszen így lehet megítélni, hogy a terminusjelölt-lista mennyi nem odatartozó elemet tartalmaz, illetve mennyi terminus hiányzik onnan.

Ezután bemutattunk különböző terminológiakivonatolókat, amelyekre mi is építettünk az alkalmazás létrehozásakor. A mi terminológiakivonatolónk elsősorban szabályalapú, de matematikai statisztikai elemeket is tartalmaz, ezért nagy fedéssel és kisebb pontossággal dolgozik. A problémák lehetséges forrása a mintakorpusz viszonylagosan kis mérete, valamint a túlzott szabályalapúság volt, hiszen minden mintát lehetetlen felsorolni, meg

kell elégednünk csak a legfontosabbakkal; ezen kívül problémát okozott még a pontosság növelése végett beépített bővítmények közül egy-két elem, mert azok időnként éppen a terminusokat szűrték ki.

Azonban a pontosság és a fedés növelése érdekében az alkalmazást különböző bővítményekkel láttuk el. Használtunk újraíró szabályokat a terminusvariánsok megkeresésére, és referenciaszótárt a gyakran előforduló, de terminusnak semmiképpen sem tekinthető elemek kiszűrésére. Ezen kívül alkalmaztunk még biztosan terminusokat tartalmazó listát is (ezek főleg idegen nyelvű szavak), és egy olyan listát, amely azon elemeket tartalmazza, amelyeket biztosan terminus követ. A többi, és főleg a terminusokra nem jellemző belső morfoszintaktikai szerkezet esetén pedig az előfordulás gyakorisága határozta meg, hogy az adott elem terminusjelölt-e vagy sem. Ezen kívül figyelembe vettük még az eredeti szöveg tipográfiai jegyeit is.

FORRÁS

Eckel, Bruce 2004: *Penser en Java* (2. kiadás). [http://penserenjava.free.fr/]

HIVATKOZÁSOK

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.

Bourrigault, D. 1994: *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir des textes.* Doktori értekezés. Paris: École des Hautes Etudes en Sciences Sociales.

Cabré, M. T. 1998: *Terminology. Theory, methods and applications,* John Benjamins Pub Co.

Cabré Castellvi, M. T. – Estopa Bagot, R. – Vivaldi Palatresi, J. 2001: Automatic term detection: A review of current systems, in Bourigault, D. – Jacquemin, C. – L'homme, M-C. eds.: *Recent advantages in computational terminology,* Amsterdam – Philadelphia: John Benjamins, 53–88.

Daille, B. 1994: *Approche mixte pour l'extraction automatique de terminologie: statistique lexicale et filtres linguistiques.* Doktori értekezés. Paris: Université de Paris VII.

Enguehard, Chantal. 2005: Un banc de test pour la reconnaissance de termes en corpus, in Williams, G. ed.: *La linguistique de corpus,* Rennes, Presses Universitaires de Rennes, 273–286.

Jacquemin, C. 2001: *Spotting and discovering terms through natural language processing,* Cambridge (Mass.), MIT Press.

Kis Á. – Kis B. – Pohl G. 2004. A számítógépes terminológiakivonatolás új megközelítése, in *A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete,* Szeged: Szegedi Tudományegyetem, 63–72.

Kis B. 2005: Automatikus terminológia keresés számítógéppel – kísérlet, *Fordítástudomány* 7/1. 84–96.

Maynard, D. – Ananiadou, S. 2001: Term extraction using a similarity-based approach, in Bourigault, D. – Jacquemin, C. – L'homme, M-C. éd.: *Recent advantages in computational terminology,* Amsterdam – Philadelphia: John Benjamins, 261–278.

Meilland, J-C. – Bellot, P. 2005: Extraction automatique de terminologie à partir de libellés textuels courts, in Williams, G. éd.: *La linguistique de corpus,* Rennes, Presses Universitaires de Rennes, 357–370.

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

Sager, J.C. 2000: Pour une approche fonctionnelle de la terminologie, in Béjoint, H. – Thoiron, P. éds.: *Le sens en terminologie,* Lyon: Presses Universitaries de Lyon.

Vasconcellos, M. 2001: Terminology and Machine translation, in Wright, S-E. – Budin, G. éds.: *Handbook of terminology management* 2, 697–723.

Wüster, E. 1976: La théorie générale de la terminologie. Un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets, *Actes du colloque international de terminologie,* Québec 5-8octobre 1975, Québec.