# Iskanje pragmatičnih enot v neoznačenem korpusu: primer kažipotov

**Agnes Pisanski Peterlin**

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
agnes.pisanski@guest.arnes.si

**Povzetek**

V prispevku je predstavljen poskus uporabe korpusne analize za iskanje pragmatičnih enot v neoznačenem specializiranem korpusu 46 slovenskih poljudnoznanstvenih člankov; v korpusu so z elektronskim iskanjem identificirani kažipoti, ki so vrsta metabesedila. Iskanje generira 2782 zadetkov, od katerih jih je 8,1 % identificiranih kot kažipoti. Nadaljnje ročno iskanje identificira dodatne primere kažipotov; pokaže se, da je elektronsko iskanje uspešno identificiralo nekaj več kot 80 % vseh kažipotov. Predstavljene so možnosti izboljšav v iskanju, kot npr. redukcija seznama iskanih izrazov na tiste, ki so bili pred tem najdeni v poljudnoznanstvenih člankih, iskanje nizov besed namesto posameznih besed in razširitev seznama iskanih besed.

**Search for Pragmatic Units in an Untagged Corpus: The Signpost Case**

The paper presents an attempt to use corpus analysis to search for pragmatic units in an untagged specialized corpus of 46 Slovene popular science articles; the corpus is searched electronically for signposts, a type of metatext. The search yields 2782 hits of which 8,1% are identified as signposts. A subsequent manual search process identifies additional instances of signposting, revealing that the electronic search has successfully identified just over 80% of all the signposts. Possibilities for improving the search, such as reducing the list of search words to those previously found in popular science articles, using word combinations instead of single words and expanding the list of search words are discussed.

## 1 Uvod

Čeprav se je korpusni pristop sprva uporabljal zlasti v analizah leksiko-gramatikalne tematike, se je v zadnjih desetih letih začel razširjati tudi na področje besediloslovnih raziskav. Te so zastavljene na različne načine: Gorjanc (2005: 71) navaja, da je bilo v zadnjem času veliko energije vložene v besediloslovno označevanje, pri čemer posebej izpostavlja področje slovnične kohezije. Poleg slovnične kohezije pa se besediloslovne raziskave pogosto lotevajo tudi drugih vrst problematike, pri čemer pa enotnega in sistematičnega pristopa ni.

V nekaterih primerih se avtorji raziskav odločijo sami označiti korpus, ki so ga zbrali za raziskavo, z besediloslovnimi oznakami, ki so za njihovo raziskavo uporabne; Upton in Connor (2001) npr. ročno označita besedilo po retoričnih korakih. Po tako označenem korpusu je iskanje enostavno in učinkovito, vendar so takšen korpus in oznake običajno manj uporabne za druge raziskave. Nekateri raziskovalci poskušajo uporabljati korpusne oznake, ki označujejo nekoliko bolj splošne besediloslovne kategorije. Eden bolj posrečenih poskusov v tej smeri je RST, Teorija retorične strukture (angl. Rhetorical Structure Theory), ki sta jo osnovala Mann in Thompson (1988) in v okviru katere so nastale razmeroma splošne oznake za strukturo besedila. Primere korpusov označenih po načelih RST lahko najdemo na spletni strani RST, http://www.sfu.ca/rst. Zanimiva primera raziskav v korpusih označenih po načelih RST predstavljata Taboada (2006), ki analizira rabo diskurznih označevalcev, in Burstein in Marcu (2003), ki primerjata dva sistema avtomatične diskurzne analize, s katero identificirata tezne in zaključne stavke v študentskih esejih.

Poleg tovrstnega pristopa, ki predpostavlja, da je za kvalitetno besediloslovno korpusno raziskavo najkoristneje posebej označevati besedilne korpuse za izbrane besediloslovne kategorije, pa so danes mnoge besediloslovne raziskave, zlasti tiste, ki se osredotočajo na strokovni jezik, zasnovane na drugačnem pristopu: avtorji za potrebe svoje raziskave pripravijo lasten specializiran korpus, ki je največkrat neoznačen, nato pa si pri analizi pomagajo z obstoječimi programi za besedilno analizo in konkordance (prim. Flowerdew, 2002: 96, za opis stanja na področju angleščine v znanosti). V tem prispevku predstavljam primer tovrstne raziskave in ob dobljenih rezultatih opozarjam na nekatere težave, ki se v zvezi s takšnim iskanjem pojavljajo.

## 2 Kažipoti

Kažipoti so elementi, s katerimi tvorec besedila bralcu oziroma poslušalcu napoveduje vsebino besedila, ki sledi, ali pa se sklicuje na tisto, kar je bilo že povedano v istem besedilu, npr. *v nadaljevanju bomo pokazali, da....* ali pa *kot smo že omenili....* Kažipote uvrščamo med metabesedilne elemente; to so deli besedila, ki bralcu ali poslušalcu služijo za orientacijo po besedilu oziroma nakazujejo odnos avtorja do sprejemnika besedila, k vsebini besedila pa ne prispevajo.

### 2.1 Zakaj so pomembni?

Poleg tega, da so raziskovalcu retorične strukture zanimivi s teoretičnega vidika, kot vrsta metabesedila, ki predstavlja nadvse kompleksno, problematično in heterogeno kategorijo, so zanimivi tudi iz različnih praktičnih razlogov. Raziskave so pokazale, da se jeziki, vede in žanri v rabi metabesedilnih elementov med seboj razlikujejo (npr. Bäcklund, 1998, Dahl 2004); to je bilo pokazano tudi za slovenščino v primerjavi z angleščino in za različne žanre in stroke (Pisanski Peterlin, 2005). Glede na to, da se tovrstnih razlik pravzaprav slabo zavedamo, so možnosti napak pri oblikovanju besedila v tujem jeziku ali v žanru, ki piscu ni domač, precejšnje.

### 2.2 Zakaj so problematični za elektronsko iskanje?

Kot številne druge besediloslovne kategorije so kažipoti problematični, ker niso formalna, temveč pragmatična kategorija: identificiramo jih po njihovi vlogu v besedilu in ne po njihovi obliki. Za pravilno interpretacijo metabesedilnosti je kontekst ključnega pomena. Čeprav se pogosto pojavljajo v podobnih ali enakih oblikah (npr. *omenjeni primeri, pokazali bomo, kot že rečeno,...*), je včasih njihova oblika povsem netipična (npr. *presenečenj pa še ni konec*). Prav tako pojavitev tipične oblike še ne pomeni, da ta označuje kažipot. Besedna zveza *opisane vrste* je kažipot, če se nanaša na vrste, ki jih je avtor besedila predhodno opisal v istem besedilu (v smislu *zgoraj/prej opisane vrste*), če pa se nanaša na vrste, ki so jih različni raziskovalci opisali do sedaj (v nasprotju s takimi, ki jih še niso opisali), pa to ni kažipot, temveč del vsebine besedila.

A kljub temu, da metabesedilo ni formalna kategorije, se avtorji, ki metabesedilne elemente raziskujejo v angleščini (Hyland, 2004) in v drugih jezikih (npr. Dahl, 2004 v norveščini in francoščini), vse pogosteje odločajo za elektronsko iskanje po posameznih tipičnih izrazih, saj le tako lahko raziskujejo rabo v velikih vzorcih in so njihove ugotovitve lahko statistično veljavne. Za iskanje po korpusu nekateri avtorji uporabljajo različne obstoječe programe za besedilno analizo in konkordance, kot so WordPilot 2000 (uporabljen v Hyland, 2005 in Harwood, 2005), WordSmith Tools (uporabljen v Hewings, M. in Hewings, A., 2002), MonoConcPro (uporabljen v Hyland, 2004), drugi pa za iskanje uporabljajo program, narejen "po meri" prav za iskanje po njihovem korpusu (npr. Dahl, 2004).

Ob tem se odpira zanimivo vprašanje, kakšne so specifične značilnosti takega iskanja v slovenščini, ki je v nasprotju z angleščino morfološko mnogo bogatejša, kako uspešno je tovrstno iskanje in s katerimi težavami se pri njem srečujemo.

## 3 Korpus in metoda

### 3.1 Korpus

Analizirani korpus je bil sestavljen za potrebe te raziskave in obsega 46 slovenskih besedil. Vsa besedila so izvirni slovenski poljudnoznanstveni članki, objavljeni v reviji *Proteus* med letoma 1997 in 1999. Skupna dolžina korpusa je 341 486 besed.

### 3.2 Metoda

V okviru predpriprave na analizo sta bila sestavljena dva seznama iskanih izrazov. Seznama sta nastala na podlagi predhodne raziskave (Pisanski Peterlin, 2005), ki je med drugim obsegala ročno analizo 70 besedil in identifikacijo vseh pojavitev kažipotov v njih. Od omenjenih besedil je bilo 35 besedil slovenskih, v katerih je bilo identificiranih 223 kažipotov. Iz tega nabora kažipotov so bili v najprej izbrani vsi glagoli, razen tistih, ki so bili ocenjeni kot presplošni, da bi bili uporabni za iskanje (npr. *biti* v *namen je*). Izdelava seznama glagolov se zdi smiselna, kajti velika večina kažipotov vsebuje glagolsko obliko. Tako dobljeni seznam obsega 66 glagolov, to pa so:

*analizirati, dokazati, govoriti, ilustrirati, interpretirati, izogniti, izpeljati, izračunati, končati, lotiti, narediti, navesti, obdelati, obravnavati, ogledati, omejiti, omeniti, omenjati, opazovati, opisati, opozoriti, opraviti, opredeliti, opustiti, orisati, podpirati, pogledati, poiskati, pojasniti, pokazati, poskusiti, posvetiti, potrebovati, poudariti, povedati, povrniti, povzeti, predstaviti, pretresti, preveriti, prevladovati, prezentirati, prihraniti, prikazati, primerjati, pripeljati, privzeti, razgrinjati, razkriti, razlagati, razložiti, reči, seznaniti, skicirati, spoznati, ubrati, ugotoviti, uporabiti, ustaviti, videti, vprašati, vrniti, začenjati, zajemati, zanimati, zapisati.*

Ločeno je bil narejen popis vseh glagolskih slovničnih oblik, ki so se pojavljale v funkciji kažipotov (npr. velelnik za prvo osebo množine, kot npr. *poglejmo*, nedoločnik v povezavi z modalnim glagolom, npr. moramo *pogledati*, prva oseba množine v sedanjiku, npr. v članku *pokažemo* itd.); kjer so se pojavljale večbesedne oblike (npr. *bomo pokazali*), je bil v analizo zajet le polnopomenski glagol, torej deležnik *pokazali.* Takih oblike je bilo enajst, in sicer *nedoločnik*, *velelnik za prvo osebo množine, 1. oseba ednine v sedanjiku, 1. oseba množine v sedanjiku, 3. oseba ednine v sedanjiku, 3. oseba množine v sedanjiku, deležnik na -l, deležnik na -la, deležnik na -lo, deležnik na -li in deležnik na -le*, kar pomeni, da je potrebno za 66 glagolov generirati seznam 726 glagolskih oblik. Morfološko bogata slovenščina je v primerjavi z morfološko revno angleščino nekoliko bolj zahtevna pri predpripravi na analizo, a hkrati omogoča izločitev tistih oblik, ki se nikoli ne pojavljajo kot kažipoti (npr. druga oseba ednine), kar lahko pomeni nekoliko bolj osredotočeno iskanje z manj lažnimi zadetki. Žal v primeru kažipotov glagolskih oblik, ki bi jih lahko v celoti izločili, ni veliko, v primeru drugih metabesedilnih elementov ali drugih funkcijskih kategorij, pa je situacija nekoliko drugačna in je tovrstno sortiranje lahko zelo koristno.

Nato je bil na podlagi že omenjenega nabora kažipotov narejen še seznam drugih izrazov (samostalniki, pridevniki, deležniki na *-n/-t* in prislovi), ki bi utegnili služiti kot potencialni identifikatorji kažipotov. Kriteriji za izbor so bili pri tem seznamu mnogo ožji kot pri glagolih: pregled izvornega nabora kažipotov je pokazal, da je pomen kažipota najpogosteje vsebovan v glagolu, med tem ko večina samostalnikov in spremljevalnih besed nima nujno le metabesedilnega pomena. Vendar pa se tudi med neglagoli pojavljajo tipični deli kažipotov (npr. samostalniki, kot so *članek, namen, razdelek* itd., pridevniki, kot so *zgornji, naslednji, spodnji*,... in prislovi, kot so *najprej, tu, zgoraj*). Deležniki na *-n/-t* so v svojem izvoru seveda glagolske oblike, vendar jih po funkciji že Toporišič (2000) uvršča med pridevniško besedo. Prav zato in zaradi svojih morfoloških značilnosti so zbrani na seznamu neglagolskih oblik. Poleg tega jih. Zaradi razmeroma omejenega števila glagolov, ki so se v izvornem seznamu pojavljali v obliki deležnikov na *-n/-t*, se je zdelo smiselno takšno omejeno število deležnikov tudi obdržati. Seznam neglagolskih izrazov za iskanje kažipotov je tako obsegal naslednje besede:

- samostalnike: *analiza, faza, razprava, članek, način, namen, opis, podatek, postopek, primer, prispevek, razdelek, rezultat,*

*sestavek, začetek, zapis, zgled, dejstvo, delo, nadaljevanje, področje, poglavje, vprašanje, pot, rešitev, ugotovitev,*

- pridevnike in deležnike na -n/-t *gornji, naslednji, našteti, obravnavani, omenjani, omenjeni, opisani, prejšnji, pričujoči, uvodni, spodnji, zadnji, zgornji*, ter številnik *prvi* (po analogiji z *zadnji*)
- prislove *doslej, kmalu, najprej, nato, nazadnje, pozneje, pravkar, predhodno , prej, sedaj, spodaj, tu, tukaj, uvodoma, zdaj, zgoraj, že*

V nasprotju z glagolskimi oblikami na prvem seznamu na drugem seznamu ni mogoče izločiti morfoloških oblik samostalnikov in pridevnikov (vključno z deležniki na *-n/-t*), saj se v metabesedilni vlogi lahko pojavijo v vseh sklonskih oblikah in v vseh treh številih, pridevniki in deležniki pa tudi v vseh spolih. Tako se npr. samostalnik *članek* v vlogi kažipota sicer izrazito tipično pojavlja v kombinaciji *v+članku*, vendar ni nobenega razloga zakaj se ne bi pojavljal tudi v primerih, kot je npr. *članek predstavlja,... namen članka je..., s tem člankom želim prikazati...* Na enak način kot pri glagolih generiramo seznam neglagolskih oblik, ki obsega 401 obliko.

Analiza korpusa je bila narejena s programskim orodjem za konkordance WordSmith Tools (Scott, 1996), verzija 4.0. Sledila je ročna izločitev vseh zadetkov, ki niso imeli metabesedilnega pomena. Nato so bila vsa besedila v testnem korpusu še enkrat analizirana ročno, da bi identificirali vse primere kažipotov, ki jih elektronsko iskanje ni zajelo, s čemer bi oceni uspešnost uporabljene metode. Nazadnje so bila besedila še enkrat računalniško analizirana s programskim orodjem za konkordance Wordsmith Tools, tokrat z nekoliko modificiranim seznamom izrazov, z namenom izpopolnitve metode iskanja.

# 4 Rezultati

V **tabeli 1** so predstavljeni rezultati iskanja po obeh seznamih in skupni seštevek: v prvi vrstici je predstavljeno število zadetkov elektronskega iskanja, v drugi vrstici je navedeno število zadetkov, ki so ostali po ročni izločitvi lažnih pozitivnih zadetkov in so bili identificirani kot kažipoti, v tretji vrstici pa je razmerje med pravimi pozitivnimi zadetki in vsemi zadetki elektronskega iskanja izraženo v odstotkih.

| | Glagoli | Neglagoli | Skupaj |
|---|---|---|---|
| Vsi zadetki | 731 | 2051 | 2782 |
| Kažipoti | 54 | 172 | 226 |
| % vseh zadetkov | 7,4 % | 8,4 % | 8,1 % |

**Tabela 1: Rezultati elektronskega iskanja kažipotov in ročnega čiščenja dobljenih rezultatov**

Iskanje po seznamu glagolov je prineslo 731 zadetkov, iskanje po seznamu drugih izrazov pa 2051 zadetkov. Od tega je bilo med identificiranimi glagolskimi oblikami 54 oziroma 7,4 % takih, ki so bili v resnici v vlogi kažipota, med ostalimi oblikami pa 172 oziroma 8,4 % vseh.

Pomemben je tudi podatek o številu primerov, ki so se podvajali na obeh seznamih; takih primerov je bilo 31. Nekateri primeri so se podvajali znotraj istega seznama, taka sta bila med glagoli 2 primera, med ostalimi pa 36 primerov. Končni seznam vseh kažipotov, ki jih je elektronsko iskanje po korpusu identificiralo, tako obsega 157 primerov.

Rezultati ročne analize so pokazali, da je bilo elektronsko iskanje kažipotov delno uspešno. Ročna analiza je namreč v korpusu identificirala še dodatnih 34 primerov, kar pomeni, da je elektronsko iskanje zajelo nekaj več kot 82 % vseh pojavitev. Skupno število vseh identificiranih kažipotov je 191, kar pomeni, da se v povprečju pojavlja 4,1 kažipot na besedilo, oziroma nekaj manj kot 0,6 na 1000 besed.

# 5 Diskusija

Pri avtomatskem iskanju kažipotov v korpusu sta se pojavila dva problema: prvi je veliko število lažnih pozitivnih zadetkov (pravih pozitivnih je manj kot 10 %), drugi pa je dejstvo, da z metodo uspešno najdemo le nekaj več kot 80 % kažipotov.

## 5.1 Lažni zadetki

Velik odstotek lažnih pozitivnih zadetkov v pričujoči raziskavi ni bil posebej problematičen, saj je uporabljeni korpus razmeroma majhen, zato je bilo ročno izločanje lažnih pozitivnih obvladljiva naloga; v večjem korpusu bi bilo to bistveno težje. Hyland (2004: 136–7) vprašanje lažnih pozitivnih zadetkov rešuje s pomočjo statistike: med številnimi zadetki jih naključno izbere 50 in med njimi identificira tiste, ki se pojavljajo v metabesedilni vlogi, nato pa izračuna isto razmerje za celoten korpus. Zdi se, da je takšen pristop primeren za grobo ugotavljanje deleža metabesedilnih elementov v celotnem korpusu, čeprav 50 primerov ni velik vzorec. Vsekakor pa s statistično posplošitvijo ne moremo ugotavljati subtilnih razlik v dveh podobnih korpusih, saj so raziskave pokazale, da so takšne razlike lahko razmeroma majhne. Prav tako podatki pridobljeni s statističnim preračunavanjem niso najbolj primerni za nadaljnje ugotavljanje drugih lastnosti kažipotov, npr. lokacije, referenčnega dosega, referenčne razdalje itd.

### 5.1.1 Krajšanje seznama

Ena od možnosti za izboljšavo iskanja je sprememba oziroma krajšanje seznama iskanih izrazov. Seznam bi lahko skrajšali v številu slovničnih oblik ali leksikalnih enot. Pregled ustreznosti nabora glagolskih slovničnih oblik pokaže, da od enajstih oblik tri niso generirale pravih pozitivnih zadetkov; te oblike so prva oseba ednine v sedanjiku ter deležnika na *-lo* in *-le*. Izpuščanje deležnikov na *-lo* in *-le* bi bilo nesmiselno, saj sta to le varianti deležnika na *-l*, ki je sicer v oblikah na *-l, -la* in *-li* generiral razmeroma velik delež pravih pozitivnih zadetkov in bi se v drugem naboru besedil lahko pojavljale tudi oblike na *-lo* in *-le*. Izpuščanje prve osebe ednine v sedanjiku prav tako ni smiselno: čeprav je očitno, da se v danem korpusu tudi v člankih, kjer je avtor en sam mnogo pogosteje pojavlja prva oseba množine, pa vendarle naletimo tudi primer kažipota v prvi osebi ednine v prihodniku, npr. *bom opisal*.

Pregled ustreznosti nabora leksikalnih enot daje boljše možnosti za krajšanje. Izhodiščna predpostavka pri

izdelavi seznama je bila, da je, glede na to, da so kažipoti pragmatična in ne formalna kategorija, smiselno na seznam uvrstiti čim več različnih izrazov, ki se v tej vlogi lahko pojavljajo. Podroben pregled zadetkov to predpostavko zanika: pokaže se, da veliko večino pravih pozitivnih zadetkov generirajo eni in isti izrazi, drugi pa sistematično nastopajo v nemetabesedilni vlogi.

Eden od možnih razlogov za to, da so se na seznamu pojavili nekateri izrazi, ki so generirali same nemetabesedile zadetke je morda v izhodiščnem naboru kažipotov. Ta je bil narejen na podlagi ročne analize 35 besedil iz treh različnih žanrov: znanstvenega članka, poljudnoznanstvenega članka in univerzitetnega učbenika. Predhodne raziskave so sicer pokazale, da se žanri med seboj močno razlikujejo v rabi metabesedilnih elementov (Crismore in Farnsworth, 1990 za angleščino, Bäcklund, 1998 za angleščino, nemščino in švedščino, in Pisanski Peterlin, 2005 za angleščino in slovenščino), z vprašanjem oblik v posameznih žanrih pa se ni ukvarjala nobena od njih. Primerjava lažnih in pravih zadetkov pokaže, da je številne lažne pozitivne zadetke mogoče pripisati dejstvu, da so bili na izvorni seznam iskanih izrazov uvrščeni iz univerzitetnega učbenika ali iz znanstvenega članka.

Primer izraza, ki generira številne lažne pozitivne zadetke je *delo*. Ta podobno kot *poglavje* izhaja iz kažipotov, ki se pojavljajo v univerzitetnih učbenikih, v poljudnoznanstvenih člankih pa ga pravzaprav ne pričakujemo, čeprav v resnici lahko naletimo na primere, v katerih avtorji *razdelke* v članku poimenujejo *poglavja* (npr. *kot smo pokazali v prejšnjem poglavju*). Podobno je tudi razmeroma neverjetno, čeprav ne nemogoče, da bi avtor članek poimenoval *delo,* (npr. *v tem delu bomo pokazali*). Pregled lažnih pozitivnih zadetkov pa pokaže, da se v veliki večini pojavljajo različne sklonske oblike samostalnika *del*, ki se prekrivajo s samostalnikom *delo* (*dela, delu, del* ipd.), samostalnik *del* pa je v poljudnoznanstvenih člankih razmeroma pogost.

Druga vrsta primerov lažnih pozitivnih zadetkov izhaja iz primerov glagolov, ki se pojavljajo v kažipotih v znanstvenih člankih. Kažipoti kot so *v članku bomo analizirali, zgoraj smo dokazali, v naslednjem razdelku bomo izračunali, izpeljali smo,...* se v naravoslovnih znanstvenih člankih pogosto pojavljajo. Narava poljudnoznanstvenih člankov pa je drugačna: v teh besedilih avtorji tipično ne analizirajo, dokazujejo, izračunavajo ali izpeljujejo itd., temveč predstavljajo, povzemajo, poročajo. Zelo pogosto pa poročajo o raziskavah ali analizah drugih raziskovalcev, npr. *na univerzi v Cambridgeu so to analizirali, dokazali, izračunali, izpeljali,...* Iskanje po izrazih kot so *analizirali, dokazali, izračunali, izpeljali* tako seveda pripelje do številnih lažnih pozitivnih zadetkov.

Kot možna izboljšava seznama se torej ponuja krajšanje seznama na zgolj tiste izraze, ki so bili identificirani v poljudnoznanstvenih člankih, brez dvoma pa to pomeni tudi izgubo nekaterih pravih pozitivnih zadetkov. Novi seznam je precej krajši in obsega 31 izrazov, to pa so:

- glagoli *izračunati, narediti, navesti, omeniti, omenjati, opredeliti, pogledati, povedati, prikazati, razgrinjati, seznaniti, spoznati, ugotoviti, uporabiti, ustaviti,*
- samostalniki *članek, način, področje, razdelek, vprašanje,*
- pridevniki in deležniki *našteti, omenjeni, opisani, prejšnji, uvodni, zgornji,*
- prislovi *najprej, prej, tu, tukaj, zgoraj, že.*

Tudi v tem primeru so za vsak izraz generirane ustrezne slovnične oblike po enakih načelih kot zgoraj.

Ponovna analiza pokaže, da so rezultati takšnega iskanja bolj obvladljivi, vendar manj natančni. Iskanje s skupnim seznamom iskanih besed, ki vsebuje glagole in ostale izraze iz poljudnoznanstvenih člankov, da 990 zadetkov, med njimi je pravih pozitivnih 128, kar je nekaj manj kot 13 %. Med identificiranimi je bilo 29 ponovitev, kar pomeni, da je bilo zares identificiranih 99 kažipotov, torej le dobro polovico vseh kažipotov (191). V resnici torej ne moremo govoriti o posebni izboljšavi.

Toda modifikacija tega seznama ponuja nekaj smernic za nadaljnje izboljšave: med iskanimi besedami izstopa prislov *že*, ki generira 222 zadetkov, med njimi številne kažipote, večina pa je lažnih pozitivnih. Pregled kažipotov pokaže, da se *že* največkrat pojavlja v povezavi z glagoloma *opisati/omeniti* ali z njunimi deležniki, to pa pomeni, da se zadetki večinoma podvajajo. V resnici prislov *že* samostojno prispeva le k identifikaciji dveh zadetkov; če bi ga izpustili iz seznama iskanih besed, bi bili rezultati drugačni: med zadetki bi bilo 16,4 % kažipotov. Tudi pregled izgubljenih zadetkov pokaže zanimivo sliko: z dodatkom samo štirih izrazov (*primer, prvi, naslednji* in *opisati*) z vsemi njihovimi oblikami, bi zajeli še 23 pojavitev, to pa bi pomenilo, da je metoda uspešna skoraj 64 %. Ti rezultati nakazujejo možno smer za izboljšave pri iskanju: nekoliko modificiran seznam, ki bi bil vezan na žanr, bi lahko nekoliko zmanjšal število lažnih pozitivnih zadetkov.

### 5.1.2 Iskanje nizov besed

Ena od možnosti za izboljšavo iskanja se zdi iskanje po nizih besed, ki se v vlogi kažipotov pogosto pojavljajo skupaj. Dobljeni rezultati delno potrjujejo smiselnost takega iskanja: skoraj 60 % glagolov je uporabljenih v kombinacijah z nekim drugim izrazom z metabesedilnim pomenom, med ostalimi izrazi je ta odstotek nekoliko nižji, a vedno dosega skoraj 40 % K temu dodamo še možnost iskanja nizov besed v kateri ima le ena metabesedilni pomen, druge pa so njene slovnične sopojavnice (npr. predlogi, pomožni glagoli,...). Namesto iskanja za obliko *članku*, bi tako iskali niz *v+članku*, ki zelo pogosto nastopa v metabesedilni vlogi, izognili pa bi se lažnim pozitivnim v smislu *njegovemu članku ne moremo očitati...* Podobno bi dosegli s kombinacijo *bomo+pokazali* namesto *pokazali*: izognili bi se vrsti primerov, kot je *pokazali so...* ipd.

Težave pri takšnem iskanju pa povzroča fleksibilnost besednega vrstnega reda v slovenščini:

*V članku bomo pokazali...*
*V tem članku bomo pokazali...*
*Pokazali bomo...*
*Pokazali pa bomo...*
*V tem prvem članku bomo sedaj pokazali...*
*Zdaj pa bomo, ne glede na vse že prej izražene pomisleke, pokazali...*

Z resno zastavljeno raziskavo najpogostejših besednih kombinacij in sopojavitev bi lahko izdelali zelo uporaben seznam iskalnih nizov, toda ob tem se postavlja vprašanje o smiselnosti takšne študije. Iskanje po besednih nizih

namreč ne more biti dokončna rešitev: nekateri kažipoti ne vsebujejo tipičnih kombinacij besed.

Drugo možnost za zmanjšanje lažnih pozitivnih zadetkov predstavlja izločanje nizov, ki se tipično pojavljajo v nemetabesedilnem pomenu. Dahl (2004: 1816) navaja uporabnost avtomatičnega izključevanja kombinacij besed kot so *they+glagol* ali *he+glagol* (ta v nasprotju z *I/we+glagol* navadno ne nastopa v vlogi kažipotov) v angleščini, francoščini in norveščini. V slovenščini takšno avtomatično izločanje ni mogoče: ker osebka ni treba izraziti: namesto *in his paper, he* ***shows*** se pojavi *v članku* ***pokaže***, oblika *pokaže* pa je seveda lahko kažipot (v smislu *rezultat* ***pokaže*** *naslednja odstopanja*).

### 5.2 Izgubljeni kažipoti

Čeprav so lažni pozitivni zadetki problematični zaradi zamudnega ročnega čiščenja, pa ne ogrožajo točnosti dobljenega rezultata. Nasprotno izgubljeni kažipoti, torej tisti, ki jih elektronsko iskanje ne identificira, rezultat izkrivljajo. Zanimivo je, da se avtorji, ki opisujejo elektronsko iskanje tistih vrst metabesedilnih elementov, ki so izrazito neformalni in pragmatični (npr. Hyland, 2004, Dahl, 2004) in za katere pričakujemo, da jih elektronsko iskanje ne bo v celoti zajelo, ne ukvarjajo z vprašanjem izpuščenih pojavitev.

Kot rečeno, je bilo v pričujoči raziskavi pri elektronskem iskanju nekaj manj kot 20 % kažipotov izgubljenih. Pregled teh primerov pokaže, zakaj jih elektronske iskanje ni zajelo. Nekateri razlogi za izgubo posameznih primerov so zelo banalni in jih lahko takoj odpravimo (npr. napačen zapis besede, *ommenjena*), vendar je tovrstnih primerov malo. Za del izgubljenih kažipotov ugotovimo, da so tako netipičnih oblik in tako močno vezani na kontekst, da jih ne bi bilo smiselno predvideti za elektronski iskanje. Navajam nekaj primerov:

*.... so zelo primerne za ponazoritev*

*Poleg te podobnosti... je še nekaj, česar ne smemo prezreti.*

*...prav tako zasluži kratko predstavitev*

V nekaterih drugih primerih lahko identificiramo izraze, s katerimi bi lahko kažipot našli tudi elektronsko, vendar so ti zelo pogosti in v veliki večini primerov rabljeni nemetabesedilno, kar pomeni, da jih verjetno ne bi bilo smiselno vključevati med iskalne izraze. Tak je zlasti kazalni zaimek *ta*, ki se običajno navezuje neposredno na predhodni stavek, kar pomeni, da ni pravi kažipot. V nekaterih primerih pa je vendarle uporabljen kot kažipot (npr. *te metode...* v smislu v *prejšnjem odstavku omenjene metode*). Podobna težava se pojavlja s števniki (glavnimi in vrstilnimi): v besedilih so zelo pogosti kot kažipoti pa razmeroma redki. Če torej izključimo netipične kažipote, primerke s kazalnimi zaimkom in s števniki, ugotovimo, da v uporabljenem korpusu približno polovice izgubljenih kažipotov z elektronskim iskanjem ne bi našli.

Drugo polovico izgubljenih kažipotov bi v danem korpusu lahko našli z izboljšanim seznamom iskanih izrazov. (V konkretnem primeru bi dopolnitev z izrazi, kot so velelnik *spomnimo se,* deležniki *opazovan, prikazan, sledeč, naveden, povedano* in prislov *zdaj* uspešnost iskanja približala 90 %). Jasno pa je, da to ne pomeni, da je seznam izboljšan do največje možne mere: šele večje število novih korpusnih raziskav in ročnih preverjanj v istem žanru bi pokazalo, kakšen je optimalen seznam za iskanje, to pa za raziskavo omejenega obsega ne pride v poštev.

## 6 Sklep

Iskanje kažipotov, metabesedilnih elementov in drugih pragmatičnih enot v elektronskem korpusu omogoča obdelavo mnogo večjih besedilnih zbirk kot ročno iskanje, zato so dobljeni rezultati statistično bolj relevantni. Toda ob njih se postavlja vprašanje veljavnosti: v pričujoči raziskave je bilo pri iskanju izgubljenih skoraj 20 % pojavitev. S tem vprašanjem se sorodne raziskave (npr. Hyland, 2004, Dahl, 2004) ne obremenjujejo. Obenem se pojavi problem številnih lažnih pozitivnih zadetkov: iskanje z daljšim seznamom iskanih izrazov je generiralo seznam primerov ,v katerem več kot 90 % zadetkov ni bilo v vlogi kažipotov. Pregled razlogov za take rezultate privede do nekaterih idej za izboljšave: redukcija seznama besed na tiste, ki so se v prejšnji raziskavi (Pisanski Peterlin, 2005) pojavile v poljudnoznanstvenih besedilih, se izkaže za preveč radikalen poseg, saj se z njim močno zmanjša uspešnost iskanja, vseeno pa nakazuje smer, v kateri bi bilo mogoče delovati za optimizacijo seznama iskanih besed. Iskanje nizov besed namesto posameznih besed bi prav tako generiralo natančnejši seznam, vendar bi povečalo delež izgubljenih kažipotov. Najboljša rešitev se zdi kombinacija vseh treh pristopov: izboljšanje seznama, krajšanje seznama v povezavi z žanrom in uporaba iskanja po nizih besed za avtomatično iskanje pravih pozitivnih zadetkov med vsemi zadetki.

Iskanje v neoznačenem korpusu je v slovenščini bolj zapleteno kot v angleščini: kjer bi v angleščini iskali s seznamom 400 oblik, je potrebno za iskanje v slovenščini izdelati seznam več kot 1000 oblik. Hkrati pa izdelava tega seznama omogoča izpuščanje tistih oblik, za katere je mogoče domnevati, da se ne pojavljajo v kažipotih in je tozadevno lahko nekoliko bolj natančno kot v angleščini. Po drugi strani pa je možnost avtomatičnega izključevanja lažnih pozitivnih zadetkov na podlagi tipičnih kombinacij, kot jo za angleščino, francoščino in norveščino uporablja Dahl (2004: 1816), v slovenščini zelo omejena.

Dobljeni rezultati pokažejo, da je iskanje pragmatičnih enot, kot so kažipoti, mogoče tudi v neoznačenem korpusu v slovenščini z obstoječim programskim orodjem za besedilno analizo in izdelavo konkordanc; pri tem pa sta predpriprave na analizo in obdelave podatkov nekoliko kompleksnejši kot v angleščini. Kot alternativa se seveda ponuja tudi možnost avtomatičnega iskanja v posebej označenem korpusu in s posebej narejenim programom.

## 7 Literatura


J. Burstein in D. Marcu. 2003. A Machine Learning Approach for Identification of Thesis and Conclusion Statements in Student Essays. *Computers and the Humanities*, 37: 455–467.

I. Bäcklund. 1998. Metatext in professional writing. A contrastive study of English, German and Swedish. *TEFA* 25: 1–42.

A. Crismore, and R. Farnsworth. 1990. Metadiscourse in popular and professional science discourse. V W. Nash

ur. *The Writing Scholar.* (118–136). Newbury Park, CA: Sage.
T. Dahl. 2004. Textual metadiscourse in research articles: A marker of national culture or of academic discipline? *Journal of Pragmatics,* 36: 1807–1825.
V. Gorjanc. 2005. *Uvod v korpusno jezikoslovje.* Domžale: Izolit.
L. Flowerdew. 2002. Corpus-Based Analyses in EAP. V J. Flowerdew, ur. *Academic Discourse.* (95–114) Harlow: Longman.
N. Harwood. 'Nowhere has anyone attempted... In this article I aim to do just that': A corpus-based study of self-promotional I and we in academic writing across four disciplines. *Journal of Pragmatics*, 7: 1207–1231.
M. Hewings in A. Hewings: "It is interesting to note that..." a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, 21: 367–383.
K. Hyland. 2004. Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing,* 13: 133–151.
K. Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7 (2) 173–192.
W. Mann in S. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text,* 8 (3): 243-281
A. Pisanski Peterlin. 2005. *Konvencije rabe metabesedilnih elementov.* Doktorska disertacija. Filozofska fakulteta, Ljubljana.
M. Scott. 1996. *WordSmith Tools*. Oxford: Oxford English Software.
M. Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38: 567–592.
J. Toporišič. 2000. *Slovenska slovnica.* (4. razširjena izdaja.) Maribor: Obzorja.
T. Upton in U. Connor. 2001. Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes* 20, 313–329.