# Prepoznavanje krajšav v besedilih

*Mojca Kompara*

V prispevku so predstavljene krajšave, vsesplošno rastoči fenomen, ki je prisoten v vseh jezikih. V slovenskem prostoru so se s krajšavami ukvarjali številni slovničarji, podrobno jih je klasificiral Matej Rode, zadnja obsežna klasifikacija pa je v *Slovenskem pravopisu* iz leta 2001. Krajšave so zajete v splošnih, specializiranih, eno- in dvojezičnih slovarjih, zaradi hitre dinamike nastajanja so težje ulovljive, slovarji v pisni obliki pa izhajajo preredko, da bi jih ažurno beležili v same slovarje. V zadnjih letih nastaja vse več spletnih zbirk, ki so prosto dostopne in ažurne ter uporabniku omogočajo tudi vnos novih krajšav. Zbirke temeljijo na besedilnih korpusih ali spletnih virih in se s pomočjo pravil in algoritmov lahko oblikujejo tudi samodejno. V nadaljevanju prispevka so predstavljeni pristopi samodejnega prepoznavanja krajšav v zbirki ADAM ter metoda Sateve in Nikolova. Ob pomoči slednjega sledi primer algoritma za prepoznavanje krajšav v slovenskih besedilih.

**Identifying Abbreviations in Texts**

This article discusses abbreviations, a generally growing phenomenon present in all languages. Many Slovenian grammarians have dealt with abbreviations, they have been classified in detail by Matej Rode, and the last comprehensive classification appeared in the *Slovenski pravopis* (Slovenian Normative Guide) of 2001. Abbreviations are found in general, specialized, monolingual, and bilingual dictionaries. Because they arise quickly, they are difficult to collect, and printed dictionaries are published too infrequently to allow the dictionaries themselves be updated regularly. In recent years an increasing number of online databases have appeared; these are freely accessible and updatable, and also allow users to enter new abbreviations. The databases are based on lexical corpora or online sources, and can also be automatically formatted using rules and algorithms. The article continues by presenting automatic abbreviation-recognition procedures in the ADAM databases as well as the Satev-Nikolov method. With the help of this method, an example is given of an algorithm for recognizing abbreviations in Slovenian texts.

## 1 Uvod

V prid ugotovitvi, da so s krajšavami križi in težave (Gabrovšek idr. 1994: 164), je več argumentov. Krajšav je res zelo veliko, nastajajo tako rekoč vsakodnevno in krajšavni slovarji jim le s težavo sledijo. Prispevek predstavlja sodobne načine

beleženja krajšav, predstavi algoritme, ki v besedilu prepoznajo krajšavo in krajšavno razvezavo. V prispevku so zastavljena vprašanja, kako so algoritmi zasnovani, kako delujejo in koliko so učinkoviti. V nadaljevanju je predstavljen poskus algoritma za prepoznavanje krajšav v besedilih in prikaz izsledkov.

## 2 Krajšave in kratice

Za potrebe prepoznavanja krajšav v besedilih je treba krajšave najprej opredeliti. Velikokrat se pojavi vprašanje, kaj pravzaprav so kratice, akronimi in krajšave, v čem se razlikujejo, kaj jih združuje in kako so definirani v različnih priročnikih. Pri rabi pomenov terminov, kot so npr. kratice, akronimi in krajšave, se prispevek naslanja na Toporišičevo *Enciklopedijo slovenskega jezika* (1992). *Kratica* je definirana kot beseda, nastala iz sklopljenih krnov večbesedne zveze, npr. *OF* iz *O + F*. Sčasoma se občutek za kratičnost lahko izgubi (prim. *laser*). Sopomenka kratice je akronim. *Krajšava* je definirana kot vse, kar je okrajšano, tudi simbolno ali kratično, npr. *nam.* (*namesto*).

## 3 Zgodovinski oris obravnavanja krajšav

Kljub temu da so krajšave v porastu predvsem v zadnjih letih, le niso novodobni pojav. Razlaga kratic je bila zapisana že v Levčevem *Slovenskem pravopisu* iz leta 1899, nato pa še v Breznikovem iz leta 1920 in v Beznikovem in Ramovševem iz leta 1935. Kratice je mogoče najti tudi v pravopisih iz let 1950 in 1962 (Logar 2003: 131). V pravopisih iz let 1950 in 1962 je kratica nadpomenka za druge tipe krajšav. V *Slovenskem pravopisu* iz leta 2001 (SP 2001) ostaja krovni pojem krajšava. Poimenovanja so se v preteklosti torej izmenjevala.

Krajšave so definirali tudi številni slovničarji, npr. Peter Dajnko v slovnici *Lehrbuch der Windischen Sprache* iz leta 1824, Anton Janežič v delu *Slovenska slovnica s kratkim pregledom slovenskega slovstva ter z malim ciriliškim berilom za Slovence* iz leta 1854, Anton Breznik v delu *Slovenska slovnica za srednje šole* iz leta 1916 (Kompara 2005: 14–20).

Tudi Tomo Korošec je v članku O krajšavah (Korošec 1993) natančno analiziral krajšave in se pri delu opiral predvsem na pravopisna pravila iz leta 1990 (SP 1990) in na Toporišičevo *Enciklopedijo slovenskega jezika* (Toporišič 1992). SP 1990 ima poglavje z naslovom *Krajšave* in ločuje med kraticami, formulami, simboli in okrajšavami. Vse te izraze združuje pod nadpomenko krajšava. Pojem krajšave se ne zamenjuje več z vrstnimi poimenovanji. SP 1990 vsebuje še *Slovarček manj znanih jezikoslovnih izrazov*, ki predstavlja definicije krajšave, kratičnega imena, krna in okrajšave (Korošec 1993: 15–27). Obsežen nabor slovenskih krajšav lahko najdemo tudi v samem slovarskem delu *Rečnika jugoslovenskih skraćenica* (Zidar 1971), prisotne so tudi v slovarskem delu *Slovarja tujk* (Verbinc 1968), nekaj jih je mogoče najti tudi v dodatku *Slovarčka tujk in kratic* (Verbinc 1969).

### 3.1 Klasifikacija krajšav po Rodetu

Prvi podrobnejši poskus opredelitve krajšav je podal Rode leta 1974. Deli jih na krajšave, okrajšave in kratice. Okrajšave se pišejo z malimi črkami in za njimi stoji pika, ko jih preberemo, pa jih vedno izgovorimo celotno, tj. neokrajšano, npr. *l.* beremo kot *leto.* Kratice se pišejo z velikimi črkami in brez pik in jih pri branju ne razvezujemo, torej ne izgovorimo celega neokrajšanega besedila. Med krajšave ne sodijo znamenja, simboli in okrajšanke. Dogovorjeni znaki so znamenja in simboli, z njimi pa zaznamujemo pojme, ne da bi izpostavili celotno ime pojma. Ko jih preberemo, izgovorimo ime pojma in ne znaka, npr. *kg* beremo kot *kilogram* in ne [kage]. Okrajšanke spadajo med samostojne besede, ki imajo lastno vsebino in obliko, nastale pa so iz kratic po besedotvornih pravilih in imajo lastno podstavo, iz katere tvorimo nove besede, npr. *Skoj*, *skojevec*, *skojevka*, *skojevski*. (Rode 1974: 215)

#### 3.1.1 Okrajšave

»Okrajšave so najstarejša oblika krajšav« (Rode 1974: 215). So priložnostne in ustaljene. Priložnostne si zamisli pisec in jih uporablja zgolj v določenem kontekstu. »Ko se raba takih, priložnostnih okrajšav razširi in jo sprejme večji krog uporabnikov, postanejo ustaljene« (Rode 1974: 215). Ustaljene okrajšave naj bi znal prebrati vsak povprečen bralec. V jezikovnem sistemu pa so ustaljene le okrajšave. Po obliki se okrajšave delijo na nesestavljene in sestavljene. Gre za proces, po katerem prvotno, neokrajšano besedilo tvori ena beseda, lahko tudi dve in več (Rode 1974: 215).

##### 3.1.1.1 Nesestavljene okrajšave

Nesestavljene okrajšave nastanejo s prekinitvijo ali krčenjem. Pri prvem postopku, tj. pri prekinitvi, se zapiše le začetni del neokrajšanega besedila, npr. *o*. za ‘oče’ ali *tel*. za ‘telefon’. Pri krčenju pa se zapišejo le določeni deli neokrajšanega besedila, npr. *dr.* za ‘doktor’ ali ‘doktorica’, *ga.* za ‘gospa’, *rkp.* za ‘rokopis’ (Rode 1974: 216).

##### 3.1.1.2 Sestavljene okrajšave

»Sestavljene okrajšave delimo v dve skupini: take, katerih člene pišemo ločeno, in take, katerih člene pišemo skupaj« (Rode 1974: 216). »Prve so lahko sestavljene iz dveh členov, npr. *d. d.*, *ing. arh.* ali iz treh in več, npr. *dr. h. c.*, *z. z o. z.*« (Rode 1974: 216). Okrajšave, kot so *npr.*, *itd.* in *ipd.*, se pišejo skupaj (Rode 1974: 216).

#### 3.1.2 Kratice

Glede na način nastanka so kratice »inicialne, zlogovne in kombinirane« (Rode 1974: 216).

##### 3.1.2.1 Inicialne kratice

Inicialne kratice so sestavljene iz začetnic, tvorimo jih tako, »[…] da vsaki besedi prvotnega neokrajšanega besedila vzamemo prvo črko in jih združimo (*R*epubliška *i*zobraževalna *s*kupnost *RIS*)« (Rode 1974: 216). Glede na to, kako se preberejo, jih je mogoče deliti na glasovne in črkovalne. Pri glasovnih je mogoče brati kratico

»kot besedo, ki jo tvorijo inicialke: *NUK [nuk]*, *VOS [vos]*« (Rode 1974: 216). »Črkovalne kratice tvorimo tako, da v besedo družimo imena inicialk« (Rode 1974: 216). Lahko se berejo po domače ali tuje. »Po domače jih izgovarjamo ali poimensko ali s polglasnikom. Poimensko beremo kratice tako, da družimo v besedo imena posameznih inicialk, kot jih poznamo pri abecedi: *UKV [u-ka-ve]*, *PTT [pe-te-te]*. S polglasnikom izgovarjamo kratico tako, da vsakemu soglasniku dodamo še polglasnik. [...] Po tuje izgovorimo črkovalne kratice tako, da v besede družimo inicialke, izgovorjene tako, kot to zahteva jezik, iz katerega smo kratico prevzeli: *HJ [ha-jot]*, *BBC [bi-bi-si]*, *BCG [be-se-že]*.« (Rode 1974: 216)

#### 3.1.2.2 Zlogovne kratice

»Zlogovne kratice tvorimo tako, da vsaki besedi neokrajšanega besedila vzamemo prvi zlog ter jih združimo v novo besedo: *Na*rodni *ma*gazin – *NAMA*, *To*varna *sa*nitetskega *ma*teriala – *TOSAMA*« (Rode 1974: 217).

#### 3.1.2.3 Kombinirane kratice

»Kombinirane kratice nastanejo s kombinacijami glasovno branih črk in zlogov: *I*ndustrija *m*etalnih *pol*izdelkov – *IMPOL*, *To*varna *mo*torjev *S*ežana – *TOMOS*, *In*dustrija *ko*vinske *o*preme – *INKOP*« (Rode 1974: 217).

### 3.1.3 Znamenja in simboli

Znamenja in simboli so znaki, ki so mednarodno dogovorjeni in se v jeziku izgovarjajo skladno z zakoni jezika. »Oblikoslovno jih prilagodimo sobesedju«, npr. *1 l* beremo ‘en liter’ itd. (Rode 1974: 218). Znamenja in simboli se pišejo s črkami, s številkami, s posebnimi znamenji in s kombinacijami teh načinov, vedno pa se pišejo brez pike. »S črkami pišemo večino mednarodnih fizikalnih enot, simbole za kemične prvine, oznake na avtomobilih in podobno« (Rode 1974: 218). Kot znake lahko uporabljamo črke tudi v geometriji, zemljepisu in drugod. Velikokrat pa se poleg črk naše abecede uporabljajo tudi črke grške abecede in črke drugih črkopisov. Od okrajšav se razlikujejo po tem, da se pišejo brez pike. Med posebne znake sodijo vsa znamenja, s katerimi nadomestimo besede za določene pojme, ko nekaj napišemo. Tak primer so: *%*, +, *$* ... »Znamenja pogosto nastopajo kot kombinacija črk, števk in posebnih znakov: *– 12° C*, *20 kW*, *7,7 %*« (Rode 1974: 218). Znamenja in simbole moramo pisati v skladu z mednarodnimi dogovori. Če z znamenji tvorimo zloženke, jih pišemo z vezajem, npr.: *20-letnica*, *H-bomba*, *A-drog* (Rode 1974: 218).

## 4 Krajšave v SP 2001 in SSKJ

### 4.1 Krajšave po Slovenskem pravopisu 2001

Zadnja izčrpna klasifikacija krajšav je v SP 2001. Tudi tu je nadpomenka krajšava, deli pa se na kratico (*SAZU*), formulo (*NaCl*), simbol (*Na*, *cm*) in okrajšavo (*oz.*, *npr.*, *d. d.*) Po SP 2001 je kratica samostalnik, nastal iz začetnih delov večbesednega poimenovanja. Kratice se izoblikujejo iz besed in stalnih besednih zvez s krnitvijo

podstavnih besed in sklapljanjem. Enaka razlaga tvorbe kratic je tudi v Toporišičevi *Slovenski slovnici* iz leta 1991. Slovenščina ima vse več domačih in privzetih kratičnih in simbolnih poimenovanj, ki se izoblikujejo s krnitvijo besed ali stalnih besednih zvez. Okrnijo se navadno do začetnih črk. Take so *meter – m* in *tempus – t.* Krni iz besednih zvez se nato strnejo v kratico in npr. iz *S*, *A*, *Z*, *U* nastane kratica *SAZU*. Lahko se jih strne tudi v formulo in dobi npr. *NaCl* za *natrijev klorid*. Če je krn en sam, se ga uporablja kot simbol in ne kot kratico, recimo *H* za *vodik* ali *t* za *čas*. Simbole za merske enote se piše s presledkom in vedno za številko, recimo *35 m* 'metrov', *5 a* 'pet arov'. Krne v kraticah se lahko piše z velikimi začetnicami ali mešano, recimo: *TAM*, *SAZU*, *BiH.* Kratice se berejo kakor druge besede, recimo *TAM* [tám], ali črkovalno, recimo *SP* [espé], samo izjemoma pa se jih izgovarja, kot zahteva tuji jezik, recimo *BBC* [bibisí]. Formule se navadno bere črkovalno *CO* [ceó], simbole pa črkovalno zgolj pri narekovanju. Simbole se razvezuje v prvotna poimenovanja, velikokrat pa se jih tudi prevaja: *Na – natrij*, *t – čas*. Za kraticami, formulami in simboli ni krajšavnih pik. Izgovorjene črkovalno brane kratice se naglašujejo le na koncu: *SP* [espé], *RTV* [ertevé]. Kratice, ki se berejo nečrkovalno, s pogosto rabo prehajajo med navadno pisano besedje in se lahko pišejo tudi takole: *sit*, *Unesco*, *Tam.* Okrajšave so po SP 2001 besede ali besedne zveze, ki so zapisane okrajšano, pika pa je znamenje okrajšanosti, recimo: *oz. – oziroma*, *t. i. – tako imenovani*, *d. d. – delniška družba*. V nasprotju s kraticami so okrajšave le pisne; ko beremo besedilo, jih običajno besedno razvezujemo, recimo *prim.* preberemo kot *primerjaj(te)* in ne kot [prím]. Le pri narekovanju običajno lahko beremo okrajšave tudi črkovalno. Okrajšave besednih zvez se pišejo s presledkom za vsako okrajšano besedo: *n. m.* za 'navedeno mesto', *red. prof.* za 'redni profesor', *d. d.* za 'delniška družba'. Zloženke okrajšamo tako, da dele, ki so okrajšani, pišemo brez presledka, npr.: *l.r.* za 'lastnoročno', *lit.zg.* za 'literarnozgodovinski'. Skupaj pa se pišejo okrajšave, kot so *itd.*, *ipd.*, *npr.*, *tj.* Po pravopisu brez vmesne pike pišemo okrajšave ene besede, iz katere jemljemo značilne črke za besedno razločevanje: *jsl.* za 'južnoslovanski', *plpf.* za 'pluskvamperfekt', *ide.* za 'indoevropski' (SP 2001: 121–122).

### 4.2 Krajšave v Slovarju slovenskega knjižnega jezika

SSKJ je enojezični razlagalni slovar, ki vsebuje okrog 110.000 gesel in podgesel in je naše temeljno slovaropisno delo. Besedni zaklad sodobnega knjižnega jezika v širšem smislu zajema prek 400.000 besed, v ožjem pa 110.000. V širšem smislu zato, ker so zajete tudi stalne besedne zveze, frazemi in termini itd. Izvzete so kratice in formule ter lastna imena (Lazar 1994). Zanimivo je dejstvo, da je trenutno edini ustrezni vir za prepoznavanje pomenov kratic in krajšav SP 2001. SSKJ razlag za kratice in krajšave (kot samostojna gesla) sploh nima, pri podstavnih besedah in zvezah pa ima prikazane (v oglatih oklepajih) kratice, npr. *a* pod *ar*, *m* pod *meter*, *dr.* pod *doktor*, pa tudi izmuzljivke kot *VOS* pod *vosovec*. Prikaz pa je žal delen, izbiren in ni sistematičen. To je za enojezični slovar nenavadno, saj tuji enojezični slovarji, npr. italijanski in angleški (ogledali si jih bomo v nadaljevanju), običajno vsebujejo tudi kratice in krajšave. Pri iskanju krajšav in krajšavnih razvezav si tako lahko pomagamo s SP 2001, ki obsega tudi precejšnje število tujih krajšav.

## 5 Krajšave v splošnih in specializiranih slovarjih

V nadaljevanju sledi pregled zastopanosti krajšav v slovarjih za enkodiranje in dekodiranje ter v tujih enojezičnih slovarjih. V ta namen je bil pregledan položaj krajšav v splošnih eno- in dvojezičnih slovarjih in specializiranih tujih slovarjih.

### 5.1 Krajšave v angleških slovarjih

V angleškem enojezičnem slovarju *Collins COBUILD English Dictionary* (Sinclair 1999) so krajšave dobro zastopane. Vključene so v sam slovar, a jim žal ni namenjeno posebno poglavje na začetku ali koncu slovarja, kot pri nekaterih drugih sorodnih slovarjih, ki so predstavljeni v nadaljevanju. Splošni dvojezični *Veliki angleško-slovenski slovar* (Grad 1998) vsebuje 100.000 gesel in na koncu kar 32 strani namenja krajšavam. *Slovensko-angleški slovar* (Grad – Leeming 1997), ki prav tako vsebuje 100.000 gesel, ima na začetku nekaj slovarskih krajšav, vendar nima posebnega poglavja, namenjenega krajšavam, niti ne vsebuje krajšav v samem slovarju. V angleščini je veliko enojezičnih krajšavnih slovarjev. Angleški specializirani slovar krajšav je na primer *The Dictionary of Acronyms and Abbreviations in Applied Linguistics and Language Learning* (*Jung* 1991) ali Penguinov *Everyman's Dictionary of Abbreviations* (Paxton 1983), ki vsebuje 27.000 gesel (Kompara 2005: 12). Nikakor ne gre spregledati nove slovarske pridobitve, *Velikega angleško-slovenskega slovarja*, ki je izšel v letih 2005–2006 v sodelovanju med založbo *Oxford University Press* in *DZS* (Gabrovšek idr. 2005–2006). Iz besedilnih korpusov *Bank of English* in *British National Corpus* ter *Fida* je bil izdelan sodoben angleško-slovenski slovar, ki spremlja stanje sodobne angleščine in slovenščine ter odseva aktualna sporazumevalna razmerja med jezikoma. Slovar celovito pokriva splošno angleško besedišče in strokovno izrazje, vključuje pa tudi britanske in ameriške različice, krajšave in lastna imena. V slovarju so številne krajšave (gre predvsem za kratice), ni pa npr. mednarodnih krajšav – kemijskih simbolov, simbolov za trdoto svinčnikov, mednarodnih avtomobilskih oznak. Slovar je izšel v dveh knjigah, krajšave so v samem slovarju, posebnega poglavja, namenjenega krajšavam, pa ni.

### 5.2 Krajšave v italijanskih slovarjih

Italijanski splošni enojezični slovar *Zingarelli* (Zingarelli 2000) ima v prilogi 24 strani krajšav. Dvojezični *Slovensko-italijanski slovar* (Kotnik 1992) ima le na začetku slovarja nekaj izključno slovarskih krajšav, v samem slovarju nima krajšav niti nima posebnega poglavja, namenjenega krajšavam. Slovar je že zelo star in vsebuje le 35.000 gesel. Novejša slovarska pridobitev, *Veliki slovensko-italijanski slovar* (Šlenc 2006), pa vsebuje 80.000 gesel in ima krajšave zajete v samem slovarju nima pa posebnega dodatka, ki bi jim bil namenjen. Na začetku slovarja je tudi seznam okrajšav, kvalifikatorjev in simbolov, ki so uporabljeni v slovarju. *Veliki italijansko-slovenski slovar* (Šlenc 1997) prav tako vsebuje 80.000 gesel in ima na začetku seznam okrajšav, kvalifikatorjev in simbolov, ki so uporabljeni v slovarju, na koncu slovarja pa je 15 strani italijanskih krajšav in slovenskih prevedkov. V italijanščini sta še dva specializirana slovarja krajšav: *Dizionario di sigle*, *abbreviazioni e simboli* (Righini 2001), ki ima 10.000 gesel, in *Dizionario delle sigle e degli acronimi* (Malossini 1999) z 8000 gesli.

### 5.3 Krajšave v nemških slovarjih

Nemški enojezični slovar *Deutsches Universal Wörterbuch* (Drosdowski 1989) ima krajšave le v samem slovarskem delu, nima pa dodatka. Prav tako imata *Veliki slovensko-nemški slovar* (Debenjak 2003) in *Veliki nemško-slovenski slovar* (Debenjak 2001) krajšave le v samem slovarju. Nemci imajo tudi krajšavni slovar *Das Wörterbuch der Abkürzungen* (Steinhauer 2005), ki šteje 50.000 nemških in tujih gesel. Žal pa tuje krajšave niso prevedene v nemščino.

### 5.4 Krajšave v španskih in francoskih slovarjih

Španski enojezični slovar *Clave: Diccionario de Uso del Español* (García Marquez 2002) vsebuje krajšave v samem slovarskem delu in ima na koncu slovarja 15 strani krajšav. V *Slovensko-španskem slovarju* (Grad 2000) ni krajšav niti v slovarskem delu niti ni dodatka, le na začetku je seznam slovarskih krajšav. *Špansko-slovenski slovar* (Grad 2001) ima slovarske krajšave tudi na začetku slovarja in jih v slovarskem delu ne vsebuje, ima pa na koncu slovarja 11 strani španskih krajšav. Francoski *Le Nouveau Petit Robert* (Robert 1996) ima krajšave v samem slovarskem delu, dodatka pa ne. *Slovensko-francoski slovar* (Jesenik 2005) ne vsebuje krajšav. *Francosko-slovenski slovar* (Grad 2004) tudi nima krajšav v slovarskem delu ima pa šest strani dodatka. Tudi Francozi in Španci imajo številne krajšavne slovarje.

Opazimo lahko, da krajšav ne vsebujejo predvsem slovarji za enkodiranje, z izjemo novega *Velikega slovensko-italijanskega slovarja* (Šlenc 2006) in *Velikega slovensko-nemškega slovarja* (Debenjak 2003). Krajšave običajno zajemajo slovarji za dekodiranje in enojezični slovarji, z izjemo SSKJ.

### 5.5 Krajšavni slovarji pri nas

Nekateri tuji enojezični slovarji vsebujejo tudi obsežne krajšavne dodatke, ki so uporabniku v veliko pomoč. V slovenščini tovrstnega novejšega slovarja še nimamo, z izjemo spletnega *Slovarčka krajšav* (http://bos.zrc-sazu.si/kratice.html), ki vsebuje 5700 slovenskih in tujih krajšav. Slovenci smo dobili prvi slovenski slovar krajšav *Kratice: mala izdaja* leta 1948 (Župančič 1948), do danes pa ni nihče napisal prenovljene izdaje ali novega tovrstnega dela. Župančičev slovar kratic je razmeroma nepoznan in redek. Slovar, ki uporablja kot krovni pojem poimenovanja *kratico*, je izšel pri Državni založbi Slovenije. V slovarju niso samo krajšave iz časa nastanka slovarja in slovenskega izrazja, temveč tudi starejše in tuje krajšave, ki se zapisujejo tako, kakor so se pojavile v knjigah, revijah in časopisih. Abecedno urejeni slovarski del zajema na 36 straneh številne krajšave, ki jih danes ne uporabljamo več oziroma niso več tako pogoste, saj so vezane na obdobje nastanka slovarja oziroma na čas pred njegovim nastankom. Nekatere krajšave, predvsem tuje, imajo poleg pomena tudi razlago. Slovarju sledijo *Dodatek*, v katerem so krajšave za učne predmete razvrščene v posebna poglavja, indijski desetiški sestav, arabske (indijske) številke, rimske številke, računski znaki, dolžinske, ploskovne, prostorninske, utežne in časovne mere ter kratice za denar in celo preračunski tečaj z nekaterimi valutami, oznake za formate papirja, znaki za strani neba in nekateri kemijski elementi (Župančič 1948).

## 6 Spletne zbirke krajšav

Ker so krajšave rastoči pojav, ki nenadoma pride v jezik in v nekaterih primerih tudi nenadoma izgine iz njega, so posledično tudi težko ulovljive. Slovarji v knjižni obliki jim po eni strani navadno ne namenjajo dovolj prostora in pozornosti, po drugi strani pa preprosto preredko izhajajo, da bi bili kos spremembam. Na voljo so še drugi načini beleženja krajšav, predvsem spletne zbirke, ki so lahko veliko bolj ažurne. Nekatere med njimi uporabniku tudi omogočajo, da sam doda krajšave in krajšavne razvezave. Primer spletne zbirke je že omenjeni *Slovarček krajšav* (http://bos.zrc-sazu.si/kratice.html) Med tujimi gesli v njem so francoske, italijanske, nemške, angleške, španske in ruske krajšave, ki se uporabljajo v slovenskem prostoru. Krajšave imajo dodan pomen, tuje pa slovenski prevod. Nekatere krajšave imajo več različnih pomenov, ki so navedeni v enem slovarskem geslu (gl. krajšavo *CD*, http://bos.zrc-sazu.si/kratice.html) in uvedeni z arabskimi številkami. Tuje krajšave imajo pred pomenom označen izvorni jezik. Slovarček sicer ni razlagalni, ampak ponekod vseeno vsebuje razlage krajšav, tujih in slovenskih (gl. krajšavi *MAT* in *CE*, http://bos.zrc-sazu.si/kratice.html), dostikrat kot pomoč pri zapletenem iskanju ustreznega pomena krajšave. *Slovarček* je nastajal od leta 2004 in je do svoje spletne postavitve postal bogatejši za več kot tisoč krajšav. Krajšave so preverjene v različnih virih (gl. seznam virov, http://bos.zrc-sazu.si/kratice.html). Prednost spletne zbirke sta možnost hitrega ažuriranja in preprosto iskanje. Podobna spletna zbirka, ki poleg področne terminologije zajema tudi krajšave, je *Evroterm* (http://www.sigov.si/evroterm/), terminološka zbirka izrazov, ki je začela nastajati med pripravljanjem slovenske različice pravnih aktov Evropske unije v okviru *Sektorja za prevajanje, redakcijo in terminologijo Službe Vlade RS za evropske zadeve* (*SVEZ*). Zbirka želi zagotoviti čim bolj poenoteno rabo terminologije pri pripravi slovenske različice pravnih aktov EU. Na spletu je na voljo od avgusta 2000, ureja in dopolnjuje se dnevno, vsebuje pa okrog 89.000 izrazov, ki izvirajo iz prevodov pravnih aktov EU in drugih dokumentov, ki jih prevajajo v državni upravi. Zbirka je večjezična in zajema številne jezike, žal pa vsi jeziki niso tako dobro zastopani kot angleščina. V zbirki so tudi krajšave, predvsem kratice angleškega izvora, ki so vedno prevedene v slovenski jezik in občasno tudi v druge tuje jezike, ni pa okrajšav in slovenskih krajšav. Zbirka vsebuje okrog 4000 krajšav. Seveda je na spletu tudi veliko tujih, predvsem angleških krajšavnih zbirk oziroma iskalnikov krajšav. Med večjimi zbirkami je *Acronym Finder* (http://www.acronymfinder.com/), ki uporabniku omogoča iskanje krajšav ali krajšavnih razvezav in tudi vnos nove definicije krajšave. Podobno deluje tudi *The Free Dictionary* (http://acronyms.thefreedictionary.com/). Spletne zbirke so nedvomno ažurnejše od knjižnih, se pa nove krajšave v jezikih pojavljajo prehitro, da bi jih ročno vnašali v obsežne zbirke. Zato se številni avtorji ukvarjajo s problematiko samodejnega prepoznavanja krajšav v besedilih in samodejnega oblikovanja spletnih slovarjev in zbirk krajšav. V nadaljevanju je predstavljenih nekaj tovrstnih pristopov.

## 7 Algoritmi za prepoznavanje krajšav

Na spletu je na voljo vedno več krajšavnih slovarjev in zbirk. Spletni slovar ima lahko zelo obsežen slovarski del, ki se preprosto in hitro ažurira in tako vključuje najnovejše krajšave. Nekateri spletni slovarji nove krajšave vključujejo kar samodejno, to jim omogočajo algoritmi za prepoznavanje krajšave in krajšavnih razvezav.

### 7.1 Zbirka ADAM

Krajšave so pomemben terminološki tip tudi na področju biomedicine. Čeprav s tega področja že obstaja nekaj zbirk krajšav, te večinoma niso namenjene širši javnosti ali pa niso dovolj izčrpne ter se osredinjajo zgolj na akronime kot tip krajšave. ADAM (http://128.248.65.210/arrowsmith_uic/adam.html) je zbirka krajšav s tega področja, ki zajema splošno uporabljene krajšave in krajšavne razvezave. Pozornost namenja akronimom in preostalim krajšavam. Rezultat raziskave je algoritem za prepoznavanje krajšav in krajšavnih razvezav iz zbirke MEDLINE (http://medline.cos.com/). ADAM je izredno natančen (97,4-odstotno) in vključuje večino pogosto uporabljenih krajšav, prisotnih v zbirki *Unified Medical Language System* (UMLS) (http://umlsks.nlm.nih.gov/) in *Stanford Abbreviation Database* (http://abbreviation.stanford.edu/). Tretjina krajšav iz ADAM-a je novih, niso bile še vključene v nobeno od omenjenih krajšavnih zbirk. 19 odstotkov novih krajšav ne sodi med akronime in zajema sedem različnih tipov krajšavno-razvezavnih parov. Zbirka ADAM je prosto dostopna.

Literatura s področja biomedicine se poveča za približno 900.000 člankov letno, kar zbirkam, kot je *Unified Medical Language System*, otežuje zapisovanje vseh krajšav. Da bi rešili težavo, so bile uporabljene številne tehnike za samodejno prepoznavanje krajšav in krajšavnih razvezav za potrebe biomedicinskih besedil, poleg tega je bilo izdelanih nekaj spletnih krajšavnih zbirk (Wren idr. 2005). Prepoznavanje krajšavnih razvezav je zelo pomembno pri reševanju pomenov krajšav pri biomedicinskih besedilih (Friedman 2000; Aronson 2001). Akronimi, ki se pojavljajo v zbirki ADAM, so besede, sestavljene iz začetnih črk ali drugih črk krajšavne razvezave, npr. *NASA* je akronim razvezave *National Aeronautics and Space Administration*. Tudi krajšava *CKB* v pomenu *brain creatine kinase* predstavlja akronim, čeprav ne sledi običajnemu zaporedju. Prisotne so tudi krajšave, ki ne sledijo določenemu leksikalnemu zaporedju, npr. krajšava *11p* v pomenu *the short arm of chromosome 11*. Opazimo da v krajšavni razvezavi ni uporabljena črka p.

V nadaljevanju je predstavljena sistematična metoda za prepoznavanje pogosto uporabljenih krajšav in krajšavnih razvezav v zbirki MEDLINE. Metoda temelji na petih zaporednih korakih: v prvem koraku izluščijo krajšave kandidatke in sobesedilo, v katerem nastopajo; v drugem koraku sledi prepoznavanje krajšavnih razvezav ob uporabi statističnih podatkov iz sobesedila; v tretjem filtriranje parov krajšav in krajšavnih razvezav glede na pravilo o dolžini; v četrtem koraku preverijo, ali so krajšave, uporabljene v besedilu, ločene od krajšavnih razvezav; v petem koraku pa skupaj združijo morfološko podobne razvezave, ki ustrezajo krajšavam (Zhou – Torvik – Samlheiser 2006: 1–6).

### 7.1.1 Koraki

V prvem koraku so luščili posamezne besede znotraj oklepajev iz zbirke MEDLINE. Da so dobili vsebino krajšav, so izluščili 3N (N pomeni dolžino krajšave kandidatke v znakih) besed levo od odprtega oklepaja v povedi. Kot primer vzemimo besedilo: »To assess the proportion of hospitalized patients who tested positive for human immunodeficiency virus (HIV) by a routine inpatient testing service [...]«. *HIV* je bil izbran kot krajšava kandidatka (krajšavne oblike) in devet (3 × 3) predhodnih besed pred odprtim oklepajem *hospitalized patients who tested positive for human immunodeficiency virus* je bilo izbranih kot sobesedilo. V primeru dvojnega oklepaja je izluščen zunanji oklepaj oziroma vsebina iz zunanjega oklepaja. Na primer, pri »decrease inserumfreetriiodothyronine (FT(3)) levels [...]«, bo *FT(3)* izbran kot kandidat. Avtorji so se odločili, da raziskujejo le krajšave znotraj oklepajev, saj je večina krajšav v besedilu definiranih kot *razvezava (krajšava)*, čeprav so tudi izjeme. Odločili so se tudi, da vključijo le samostojne besede, čeprav večbesedne krajšave seveda tudi obstajajo. Trenutni model žal še ni sposoben ločevati med večbesednimi krajšavami iz oklepajev, vključno z biomedicinskimi termini, ki niso krajšave. Da bi določili, ali je pomembno vključiti večbesedne krajšave, so preučili najpogostejše večbesedne krajšave iz zbirke *Stanford Abbreviation Database*. Opaženo je bilo, da gre pretežno za sestavljene krajšave tipa *DPP III* v pomenu *Dipeptidy1 Peptidase III*. Tovrstne primere je nujno treba obravnavati glede na enobesedne krajšave/razvezave *Dipeptidy1 Peptidase* (DPP). V zbirko ADAM so bile vključene le enobesedne krajšave, krajšave z eno črko, na primer *A–Z*, niso šteli kot pomembne za obravnavo. V zbirki *Stanford Abbreviation Database* se pogosto rabijo samo *1-adrenaline (A), 1-phosphate (P) in 1-hour (H)*. Krajšave kandidatke so bile omejene na enobesedne krajšave z dvema ali več črkovno-številskimi znaki. Izločene so bile rimske številke, ki se jih v besedilu pogosto uporablja za številčenje. Izbran je bil vzorec »razvezava (krajšava)« [f1] namesto »(razvezava) krajšava« [f2] ali »(krajšava) razvezava« [f3] ali »krajšava (razvezava)« [f4]. Preučili so naključni vzorec krajšav iz zbirke *Stanford Abbreviation Database* in opazili, da jih v MEDLINE besedilih kar 99,2 odstotka sledi vzorcu »razvezava (krajšava)«, v primeru s preostalimi možnimi pari. Da so to domnevo potrdili, so naključno izbrali tisoč parov, ki so v zbirki ADAM in niso v zbirki *Stanford Abbreviation Database*. Za vsak par so upoštevali tele štiri prej navedene oblike: f1, f2, f3 in f4. V 98 odstotkih primerov se je pojavil par *razvezava (krajšava*). Vključili so 3N-besed iz sobesedila, saj je bilo prikazano, da je pravilno razvezavo mogoče z lahkoto najti znotraj 3N-besed krajšave, ki je akronim. Rezultati so potrdili, da se to dogaja tudi s krajšavami, ki niso akronimi. Morfološko podobne krajšave so bile združene in obravnavane kot variante istega termina. Na primer *APC* se lahko zapiše kot *APC*, *Apc*, *ApC*, *aPC*, *APc*, *apc*, *AP-C* ali *Ap-C*. Po združevanju podobnih krajšav je nastala zbirka podatkov, ki pomaga pri prepoznavanju novih razvezav.

V drugem delu se ukvarjajo s težavami pri prepoznavanju krajšavnih razvezav kandidatov znotraj 3N-besed, ki stojijo levo od krajšave znotraj oklepajev. Na primer: *APC* (ali različice zapisa *Apc*, *ApC* itd.) so se v oklepajih pojavile 4579-krat v 4472 člankih. Razvezava *Adenomatous Polyposis Coli* se je pojavila 807-krat v 705 člankih na levo od krajšave (*APC*). Vprašanje je, kako prepoznati razvezave kot

ustrezno razvezavo krajšave *APC* in ne kot krajšo in daljšo razvezavo, brez uporabe kakršnih koli leksikalnih podatkov, kot je na primer sovpadanje črk. Začeli so z *APC* in preučili vsak korak. Celoten proces je formaliziran in razdeljen v naslednje zaporedje korakov:

- tokenizacija konteksta,
- štetje pojavnic v kontekstu,
- določitev kandidata razvezave,
- odstranitev neustreznih kandidatov.

Glede na dolžino so razvezave po navadi daljše od krajšav. Njihovo razmerje (dolžina razvezave/dolžina krajšave, pri čemer je dolžina določena kot število znakov) so uporabili za filtriranje parov krajšava/razvezava. 95 odstotkov enobesednih krajšav/razvezav v zbirki *Stanford Abbreviation Database* ima navedeno razmerje ≥ 2,5, zato je bila v raziskavi izbrana ta vrednost. V četrtem koraku so preverili, ali so krajšave uporabljene v besedilu ločeno od razvezav, v petem pa so združili morfološko podobne razvezave, ki ustrezajo krajšavam.

### 7.1.2 Izsledki

V zbirki MEDLINE je bilo preučenih vseh 15.433.668 navedkov (naslovov in izvlečkov). Z zgornjo metodo je bilo najdenih 512.314 parov krajšav/razvezav. Po združevanju morfoloških različic je imel ADAM 59.405 parov krajšav/razvezav. Pri merjenju kakovosti parov so ugotovili, da pri dveh naključnih izborih po tisoč parov krajšav/razvezav nastopi 23 in 29 napak; stopnja napake je 2,6-odstotna. Opazovali so tri tipe napak: npr. položaj krajšave ni bil desno od razvezave, temveč na sredini, npr. *electron (EM) microscopic examination*: *electron* je bil izluščen kot razvezava krajšave *EM* in *microscopic* je bil izpuščen. Včasih ni standardne oblike razvezave krajšave, npr. za krajšavo *CelB* je sistem beležil razvezavo *Pyrococcus furiosus*, razvezava pa je *the beta-glucosidase from the hyperthermophilic archaeon Pyrococcus furiosus*. Do tega je prišlo zaradi različnih možnosti zapisa razvezave (npr. *hyperthermostable beta-glycosidase from Pyrococcus furiosus*). V nekaj primerih se je krajšava nanašala na razvezave, ki nimajo različnih začetnih besed, ampak se končajo z enako besedo ali zaporedjem besed npr. *CCQ* je lahko *Cancer Coping Questionnaire*, *Cocaine Craving Questionnaire* ali *Common Core Questionnaire*. Nobena od navedenih oblik ni prevladovala in se ni pojavljala pogosto (Zhou – Torvik – Samlheiser 2006: 1–6).

## 7.2 Pristop Sateve in Nikolova

S prepoznavanjem krajšav sta se ukvarjala tudi Vesna Satev in Nicolas Nikolov. Obravnavala sta luščenje krajšav v srbskem jeziku ob uporabi svetovnega spleta kot korpusa. Uporaba svetovnega spleta v vlogi korpusa je razmeroma nova, vendar žal še niso na voljo ustrezna orodja za raziskovanje spleta, ki bi bila primerna za jezikoslovne potrebe. Avtorja sta izbrala avtomatično spletno preiskovanje (angl. *crawling*) kot proces zbiranja podatkov s spleta, da bi iz njega izluščila krajšave v srbskem jeziku. Pokazala sta, da se z uporabo spleta kot korpusa da najti veliko

predvsem novih krajšav. Korpusi so prinesli nove načine preverjanja jezikoslovnih hipotez, ki prej, pri ročni obdelavi podatkov, niso bili možni in predstavljivi. Sicer pa tudi korpusno raziskovanje ni brez pomanjkljivosti. Po mnenju avtorjev je lahko izdelava korpusa zelo draga, korpusi predstavljajo jezik le v določenem časovnem okviru in navadno ne ponujajo dostopa do ažurnih podatkov o jezikovni rabi ali spremembah, ki nastopajo v jeziku. To so, po mnenju avtorjev, tudi razlogi, zaradi katerih vse bolj uporabljamo splet kot vir, ki vsebuje tudi več besedilnih virov kot klasični besedilni korpusi. Splet je za nekatere jezike lahko tudi edini dostopni vir za preučevanje. Poleg tega je splet prosto dostopen skoraj kjer koli in kadar koli. V prispevku avtorja poudarjata prednosti spleta in menita, da je splet veliko boljša izbira tudi od razmeroma obsežnega korpusa (Satev – Nikolov 2008: 75).

Čeprav Chiari (2007) meni, da je splet po obsegu res lahko obravnavan kot korpus, niso pa jeziki v njem ustrezno zastopani, saj so najštevilčnejša angleška besedila, ki jih je kar 70 odstotkov, sledijo jim japonska, nemška in francoska. Splet zajema pretežno besedila o novih tehnologijah in novicah, besedila iz klepetalnic in s forumov, manj pa je regionalnih različic, primerov govora ali literarnih besedil (Chiari 2007: 54–55).

### 7.2.1 Koraki

Sateva in Nikolov sta svoje trditve podprla s primeri. Pravita, da je bilo število spletnih strani v indeksih iskalnikov leta 2005 od 10 do 20 milijard. Gre predvsem za tako imenovani »vidni del«, veliko pa je še »nevidnega dela«, to so predvsem dokumenti, ki niso dostopni prek iskalnikov. Raziskovalci skušajo odpraviti tudi te pomanjkljivosti, vendar so z delom še bolj na začetku. Jezikoslovci na spletu, ki vsebuje širok nabor besedil, lahko uporabljajo tudi spletno večjezikovnost. Splet uporablja 35,2 odstotka Angležev, 13,7 odstotka Kitajcev, 8,4 odstotka Japoncev, 9,0 odstotka Špancev, 6,9 odstotka Nemcev, 4,2 odstotka Francozov, 3,9 odstotka Korejcev, 3,8 odstotka Italijanov, 3,1 odstotka Portugalcev, 1,7 odstotka Nizozemcev in 10,1 odstotka preostalih. Po mnenju avtorjev je še ena prednost spleta kot korpusa že omenjeno dejstvo, da za nekatere jezike ostaja edini vir informacij. Jezikoslovci sicer potrebujejo čim bolj različne podatke, za odgovor na številna raziskovalna vprašanja pa zadostujejo že podatki iz standardnih korpusov, kot sta *British National Corpus* ali *Corpus of Contemporary Serbian Language*. Obstajajo pa seveda tudi primeri, ko potrebnih podatkov ne najdemo v korpusu. To se zgodi, ko je preučevani pojem redek, ko pripada tipu besedila, ki ni prisoten v korpusu ali zajema časovno omejene informacije, npr. preveč nova dejstva. Splet je kot vir informacij odličen in ugoden, saj je ažuren, prosto dostopen, kompleten, jezikovno raznovrsten, hiter in stroškovno zelo ugoden. Uporaba spleta v vlogi korpusa pa je razmeroma nova. Splet se uporablja v sodobni leksikografiji, v semantiki in predvsem pri prevajanju, poleg tega pa tudi za opazovanje sprememb v jeziku. Avtorja v nadaljevanju prikazujeta, kako se da z uporabo spleta najti nove krajšave v srbskem jeziku in potencialno tudi njihove definicije oziroma razvezave, ki jih ni mogoče dobiti s standardnimi korpusi. Avtorja navajata kot primer nekdanjo rabo besede *Kosovo* za označevanje besedne zveze *Kosovo i Metohija*. Zaradi novih trendov je zdaj namesto polnega imena v rabi krajšava *KiM*. Besede ni mogoče najti v stan-

dardnemu korpusu srbskega jezika, zato ostaja kot rešitev splet. Do informacij na spletu lahko pridemo le s pomočjo iskalnikov, tu pa se pojavi težava, saj iskalniki niso izdelani za potrebe jezikoslovcev. Zato so potrebne nove tehnike raziskav, ki temeljijo na spletnih podatkih. Na voljo je več možnosti za uporabo podatkov s spleta. Raziskovalci lahko neposredno iščejo s pomočjo iskalnika, npr. *Googla*. Čeprav iskalniki niso prirejeni za iskanje odgovorov na jezikoslovna vprašanja, jih raziskovalci vseeno uporabljajo v ta namen. Nekateri jih uporabljajo tudi za iskanje pojavnic, kar je seveda uporabno, a je lahko tudi problematično, npr. ker iskalniki ne ločujejo med velikimi in malimi črkami. Avtorja navajata kot primer pridevnik *jasna* in lastno ime *Jasna*, ki imata kot rezultat spletnega iskanja enako število zadetkov oziroma pojavitev. Podobno velja tudi za različice zapisa *white-space*, saj najdemo različne možnosti zapisa npr. *white space*, *white-space* in *whitespace*. Iskalniki, kot je denimo Google, imajo svoja pravila pri iskanju: nekatera lahko izklopimo, drugih ne. Težava je tudi dvojna pojavitev besedila, ki ga najde iskalnik, saj je lahko isto besedilo objavljeno na več različnih spletnih straneh. To pa seveda umetno zviša število pojavnic (Satev – Nikolov 2008: 75–77).

### 7.2.2 Izsledki

Po mnenju avtorjev je uporaba spleta v vlogi korpusa najboljša alternativa za analizo lastnih imen, ki so odprto vprašanje, saj nova imena nastajajo dnevno in pri tem standardni korpusi zaradi neažurnosti seveda odpovejo. Podobno je s krajšavami. Avtorja sta v raziskavo vključila akronime kot tip krajšave; naloga je bila torej najti nove akronime. Ker je korpus srbskega jezika premalo ažuren (najnovejša publikacija je iz leta 2000), sta uporabila splet kot vir. Raziskavo sta izvedla tudi na podlagi srbskega korpusa in primerjala rezultate. V srbščini so krajšave pogosto pisane z velikimi črkami, npr. *SCG – Srbija i Crna Gora* (Srbija in Črna gora) ali *SANU – Srpska akademija nauka i umetnosti* (Srbska akademija znanosti in umetnosti). Za iskanje krajšav sta pri raziskavi uporabila besede, ki niso daljše od pet črk, pri čemer sta upoštevala, da je več kot polovica črk velikih tiskanih, čeprav so v srbskem jeziku pogoste tudi krajšave tipa *BiH* in *KiM*. Izločila sta besede, ki ustrezajo zgornjim zahtevam, a se pojavljajo v sobesedilu, zapisanem z velikimi tiskanimi črkami. V nadaljevanju sta preiskovala spletne strani.

Pri tem postopku sta odstranila oznake HTML, tudi v skriptih, in nesrbska besedila. Spletni korpus, ki sta ga tako dobila, je vseboval okrog 500.000 besed. Besedilo je bilo tokenizirano, besede in sobesedilo pa zbrano v zbirki. Izluščila sta besede, ki niso bile daljše od petih črk, pri čemer je morala biti več kot polovica črk velikih tiskanih v sobesedilu, ki ni bilo zapisano z velikimi črkami. Po končanem postopku sta dobila približno 600 akronimov. 9 odstotkov akronimov se je pojavilo v obeh korpusih (v srbskem in spletnem), 33 odstotkov le na spletu, 57 odstotkov pa samo v standardnem korpusu. Avtorja opozarjata, da zaradi omejitve na pet črk niso vključene daljše krajšave, npr. *UNESCO*. Treba je tudi upoštevati, da je bil spletni korpus 50-krat manjši od srbskega korpusa, vendar sta na spletu vseeno našla nove akronime, ki jih v srbskem korpusu ni. 57 odstotkov akronimov, ki sta jih našla samo v standardnem korpusu, pojasnjujeta z dejstvom, da zajemajo imena političnih strank in organizacij ter držav, ki ne obstajajo več, npr. *SSSR*. Spletni korpus je

vključeval dnevno časopisje, ki zajema sodobne članke, in ne vsebuje zgodovinskih besedil. Avtorja menita, da se lahko delež akronimov, ki se je pojavil le na spletu, v standardnem korpusu pa ne, le še povečuje (Satev – Nikolov 2008: 77–79).

## 8 Primer algoritma za prepoznavanje krajšav v slovenskih besedilih

V nadaljevanju sta predstavljena algoritem, ki je podoben zgoraj opisanim, kot možnost samodejnega oblikovanja krajšavnih zbirk. Raziskava je bila omejena na akronime in kratice. Vir raziskave je bil dnevnik *Delo* iz leta 2007.

### 8.1 Koraki in izsledki

Najprej je bilo treba dobiti krajšavne kandidatke in kot take so bile vključene besede, ki imajo do 5 črk in so zapisane v oklepajih, npr. (*NAMA*). V dnevniku *Delo* je bilo približno 25.600 takih besed, nekatere so se pojavile tudi večkrat. Da bi prišli do želenega nabora krajšav za nadaljnje raziskovanje, je bilo treba izločiti krajšave, ki so se ponavljale, in vse krajšave, ki niso akronimi ali kratice. Izločene so bile tudi vse besede, ki niso krajšave, npr. lastna imena, pojavilo se je tudi nekaj ljubkovalnih imen, naselbinskih in nenaselbinskih imen, največ je bilo imen krajev. Za uspešno izločitev naselbinskih in nenaselbinskih imen lahko uporabimo SSKJ. Po izločitvah se je seznam zmanjšal na okrog 4000 krajšav. Da bi dobili le akronime in kratice, so bile izločene tudi vse krajšave, ki se pišejo z malimi črkami. Upoštevano je bilo tudi, da se kandidati ne pojavijo v sobesedilu, zapisanem samo z velikimi črkami. Dobljen je bil seznam, ki je štel okrog 2500 akronimov in kratic.

V vlogi kandidatk razvezav akronimov in kratic v dnevniku *Delo* je bilo opazovano sobesedilo, ki stoji levo od oklepajev, saj razvezave običajno stojijo pred krajšavo, npr. *Evropska centralna banka* (*EBC*), seveda pa ni izključena možnost razvezave na desni strani. Za prepoznavanje razvezav so bili upoštevani 4 tipi akronimov in kratic.

Prvi tip so tako imenovane prekrivne krajšave, pri katerih je število in vrsta črk enako številu besed in začetnic levo od oklepaja. Kot primer bi ponazorili kratico *FF* v pomenu *Filozofska fakulteta*. Kratica *FF* je sestavljena iz dveh velikih črk, ki sovpadata z razvezavo, sestavljeno iz dveh besed, pri čemer se obe začneta z enakima črkama kot kratica, torej s *F* in *F*. Upoštevano je bilo torej dejstvo, da se za razvezavo upošteva toliko besed, kolikor je črk v kratici, morajo pa se začeti z enakimi črkami kot kratica.

Drugi tip so kratice, ki vsebujejo predloge ali veznike, npr. *FDV* v pomenu *Fakulteta za družbene vede*, pri čemer je treba upoštevati, da sistem zajame tudi razvezave, ki imajo eno ali dve besedi več, kot je dejanskih črk v kratici, te pa sovpadajo z začetnicami v razvezavi.

Tretji tip so akronimi, ki so sestavljeni iz prvih dveh začetnih črk, npr. *NAMA* v pomenu *Narodni magazin*, pri čemer je pri razvezavi treba upoštevati prvo in drugo črko v akronimu.

Četrti tip pa so kratice s predlogi, npr. *DZU* v pomenu *Družba za upravljanje*, pri čemer predlog pri kratici ni izpuščen in nastopa tudi v razvezavi.

Po upoštevanju teh meril je bil dobljen nabor 1800 krajšavnih razvezav, ki so sovpadale s kraticami in akronimi, desno od razvezave v oklepajih. V končnem seznamu je bilo mogoče opaziti predvsem probleme s skloni: sistem je beležil vse sklone in nekatere razvezave so se pojavile večkrat. Veliko je bilo tudi tujih krajšav. Poskus pa je tudi pokazal, da se da s pomočjo preprostih navodil za prepoznavanje krajšav in krajšavnih razvezav oblikovati ustrezni nabor gesel in razvezav ter samodejno krajšavno zbirko.

## 9 Sklep

Krajšave niso novodoben pojav, saj jih je uporabljal že Cicero (Kompara 2005: 10). Postale so del našega vsakdana, nekaj, kar se je močno zasidralo v naša življenja, kar je v splošni rabi. Nastajajo dnevno, nekatere ostanejo za vedno, npr. *Nama*, druge čez čas počasi izginejo iz rabe in nanje pozabimo. Dejstvo je, da so hitro rastoči fenomen, prisoten v vsakem jeziku. Zaradi hitre dinamike jim slovarji v knjižni obliki s težavo sledijo. Tuji jeziki imajo številne krajšavne slovarje, tudi Slovenci smo bili deležni prvega slovarja krajšav, ki je izšel v letu 1948, žal pa prav zaradi hitrega nastajanja novih krajšav taki slovarji hitro zastarajo in so pomanjkljivi. Rešitev predstavlja splet, kjer je na voljo veliko terminoloških zbirk s krajšavami in krajšavnih slovarjev. Prednost spletnega slovarja je nedvomno v preprostem iskanju, poleg tega pa lahko tudi sami uporabniki ažurirajo spletne slovarje in zbirke z novimi krajšavami. Seveda je ročno beleženje krajšav zamudno, zato se je v zadnjih letih pojavila tendenca samodejnega sestavljanja krajšavnih zbirk in slovarjev. Računalniški programi s posebnimi pravili in smernicami prepoznavajo krajšave ter njihove razvezave in oblikujejo spletne slovarje, ki se ažurirajo samodejno. Nekateri temeljijo na korpusih, drugi pa uporabijo kar splet. V raziskavi, ki je bila izdelana na podlagi besedil časnika *Delo* za leto 2007, so bila uporabljena pravila za beleženje krajšav in krajšavnih razvezav. Izdelana je bila samodejna krajšavna zbirka, v kateri je 1800 krajšav sovpadalo z razvezavami. Treba je poudariti, da se številni avtorji tovrstnih algoritmov ukvarjajo izključno z akronimi in kraticami. Žal se nihče še ni preizkusil v drugih tipih krajšav. Tovrstni algoritmi nedvomno predstavljajo pot k samodejnemu oblikovanju slovarskih krajšavnih zbirk.

## Literatura

Acronym Finder [URL: http://www.acronimfinder.com].

ADAM [URL: http://128.248.65.210/arrowsmith_uic/adam.html].

Chiari 2007 = Isabella Chiari, *Introduzione alla linguistica computazionale*, Roma – Bari: Laterza, 2007, 54–55.

Debenjak 2001 = Božidar Debenjak – Doris Debenjak – Primož Debenjak, *Veliki nemško-slovenski slovar*, Ljubljana: DZS, 2001.

Debenjak 2003 = Božidar Debenjak – Doris Debenjak – Primož Debenjak, *Veliki slovensko-nemški slovar*, Ljubljana: DZS, 2003.

Drosdowski 1989 = Günther Drosdowski, *Deutsches Universalwörterbuch*, Mannheim idr.: Duden Verlag, 1989.
Evroterm: Večjezična terminološka zbirka [URL: http://evroterm.gov.si/].
Gabrovšek 1994 = Dušan Gabrovšek, Kodifikacija angleškega jezika v specializiranih enojezičnih slovarjih: Too much of everything?, *Vestnik: Društvo za tuje jezike in književnosti* 28 (1994), št. 1–2, 150–180.
Gabrovšek idr. 2005–2006 = Dušan Gabrovšek idr., *Veliki angleško-slovenski slovar Oxford*, Ljubljana: DZS, 2005–2006.
Garcia Marquez 2002 = Gabriel Garcia Marquez, *Clave: Diccionario de Uso del Español*, Madrid: Ediciones Sm., 2002
Grad 1997 = Anton Grad, *Slovensko-angleški slovar*, Ljubljana: DZS, 1997.
Grad 1998 = Anton Grad, *Veliki angleško-slovenski slovar*, Ljubljana: DZS.
Grad 2000 = Anton Grad, *Slovensko-španski slovar*, Ljubljana: DZS, 2000.
Grad 2001 = Anton Grad, *Špansko-slovenski slovar*, Ljubljana: DZS, 2001.
Grad 2004 = Anton Grad, *Francosko-slovenski slovar*, Ljubljana: DZS, 2004.
Jesenik 2005 = Viktor Jesenik, *Slovensko-francoski slovar*, Ljubljana: DZS, 2005.
Jung 1991 = Heidrun Jung – Udo O. H. Jung, *The Dictionary of Acronyms and Abbreviations in Applied Linguistics and Language Learning*, New York idr.: Peter Lang Publishing Group, 1991.
Kompara 2005 = Mojca Kompara, *Slovensko-italijanski glosar krajšav: diplomsko delo*, Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo, 2005.
Korošec 1993 = Tomo Korošec, O krajšavah, v: *XXIX. Seminar slovenskega jezika, literature in kulture*, ur. Miran Hladnik, Ljubljana: Filozofska fakulteta, 1993, 15–27.
Kotnik 1992 = Janko Kotnik, *Slovensko-italijanski slovar*, Ljubljana: DZS, 1992.
Lazar 1994 = Branka Lazar, O nastajanju Slovarja slovenskega knjižnega jezika, *Primorska srečanja* št. 160–161, 18 [tj. 19] (1994), 539–541.
Logar 2003 = Nataša Logar, Kratice in tvorjenke iz njih – aktualno poimenovalna možnost, v: *Współczesna polska i słoweńska sytuacja językowa = Sodobni jezikovni položaj na Poljskem in v Sloveniji*, ur. Stanisław Gajda – Ada Vidovič Muha, Opole: Instytut Filologii Polskej, Filozofska fakulteta, 2003,131–149.
Malossini 1999 = Andrea Malossini, *Dizionario delle sigle e degli acronimi*, Milano: Avallardi, 1999.
MEDLINE [URL: http://medline.cos.com].
Paxton 1983 = John Paxton, *Everyman's Dictionary of Abbreviations*, London: J. M. Dent & Sons, 1983.
Righini 2001 = Enrico Righini, *Dizionario di Sigle Abbreviazionie e Simboli*, Bologna: Zanichelli, 2001.
Rode 1974 = Matej Rode, Poskus klasifikacije krajšav, *Slavistična revija* 22 (1974), št. 2, 213–219.
Robert 1996 = Paul Robert, *Le Nouveau Petit Robert*, Paris: Dictionnaires le Robert, 1996.

Satev – Nikolov 2008 = Vesna Satev – Nicolas Nikolov, Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language, v: *Jezikovne tehnologije: Zbornik 11. mednarodne multikonference Informacijska družba – IS 2008*, *Zvezek C*, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Institut Jožef Stefan, 2008, 75–79.

Sinclair 1999 = John Sinclair (ur.), *Collins COBUILD English dictionary*, London: HarperCollins, 1999.

SSKJ 1– 5 = *Slovar slovenskega knjižnega jezika* 1–5, Ljubljana: DZS, 1970–1991.

*Slovarček krajšav* [URL: http://bos.zrc-sazu.si/kratice.html].

SP 1990 = *Slovenski pravopis : pravila*, Ljubljana: DZS, 1990.

SP 2001 = *Slovenski pravopis*, Ljubljana: SAZU – Založba ZRC, ZRC SAZU, 2001.

Stanford Abbreviation Database [URL: http://abbreviation.stanford.edu].

Steinhauer 2005 = Anja Steinhauer, *Das Wörterbuch der Abkürzungen*, Mannheim idr.: Duden Verlag, 2005.

Šlenc 1997 = Sergij Šlenc, *Veliki italijansko-slovenski slovar*, Ljubljana: DZS, 1997.

Šlenc 2006 = Sergij Šlenc, *Veliki slovensko-italijanski slovar*, Ljubljana: DZS, 2006.

The Free Dictionary [URL: http://acronyms.thefreedictionary.com].

Toporišič 1991 = Jože Toporišič, *Slovenska slovnica*, Maribor: Založba Obzorja, 1991.

Toporišič 1992 = Jože Toporišič, *Enciklopedija slovenskega jezika*, Ljubljana: Cankarjeva založba, 1992.

Unified Medical Language System [URL: http://umlsks.nlm.nih.gov].

Verbinc 1968 = France Verbinc, *Slovar tujk*, Ljubljana: Cankarjeva založba, 1968. S ponatisi.

Verbinc 1969 = France Verbinc, *Slovarček tujk in kratic*, Ljubljana: Prešernova družba, 1969.

Zhou – Torvik – Smalheiser 2006 = Wei Zhou – Vetle I. Torvik – Neil R. Smalheiser, ADAM: another database of abbreviations in MEDILINE [http://128.248.65.210/arrowsmith_uic/tutorial/zhou_bioinformatics_2006.pdf].

Zidar 1971 = Josip Zidar, *Rečnik jugoslovenskih skraćenica*, Beograd: Međunarodna politika, 1971.

Zingarelli 2000 = Nicola Zingarelli, *Vocabolario della lingua italiana*, Bologna: Zanichelli, 2000.

Župančič 1948 = Jože Župančič, *Kratice: mala izdaja*, Ljubljana: DZS, 1948.

JEZIKOSLOVNI ZAPISKI 15 • 2009 • 1–2

## Identifying Abbreviations in Texts

Summary

This article discusses abbreviations, a generally growing phenomenon present in all languages. Abbreviations are not a new phenomenon; they were even used by Cicero. Slovenian lexicographers have dealt with abbreviations, they were classified in detail by Matej Rode in 1974, and the last comprehensive classification appeared in the *Slovenski pravopis* (Slovenian Normative Guide) of 2001. Abbreviations are found in general, specialized, monolingual, and bilingual dictionaries. Because they arise quickly, they are difficult to collect, and printed dictionaries are published too infrequently to allow the dictionaries themselves to be updated regularly. The internet offers a good solution because it enables rapid updating and is also freely accessible. Many terminological databases with abbreviations as well as dictionaries of abbreviations are available online; the advantage of an online dictionary is unquestionably simple searching, and in addition users themselves can update online dictionaries and databases with new abbreviations. Of course, manual entry of abbreviations is time-consuming, and so in recent years there has developed a trend for automatic preparation of abbreviation databases and dictionaries such as the ADAM databases. Computer programs with special rules, methods (e.g., the Satev-Nikolov approach), and guidelines can recognize abbreviations and their equivalents written out in full and format online dictionaries that update themselves, thus increasing the entry of abbreviations. The databases are based on corpora or online sources and format themselves automatically with the help of algorithms. The presentation of an algorithm for recognizing abbreviations in Slovenian texts demonstrates that algorithms are a good solution for the future because they represent a path to automatic formatting of lexicographic abbreviation databases.