

Špela Vintar

TERMINOLOGIJA

Terminološka veda
in računalniško podprta terminografija

TERMIN
LOGIJ

Špela Vintar

TERMINOLOGIJA

Terminološka veda
in računalniško podprta terminografija

TERMINOLOGIJA. TERMINOLOŠKA VEDA IN RAČUNALNIŠKO PODPRTA TERMINOGRAFIJA.

Avtorica: dr. Špela Vintar

Recenzenta: dr. Vojko Gorjanc, dr. Apolonija Gantar

Zbirka Prevodoslovje in uporabno jezikoslovje

© Univerza v Ljubljani, Filozofska fakulteta, 2017.

Vse pravice pridržane.

Brez pisnega dovoljenja Filozofske fakultete Univerze v Ljubljani je prepovedano reproduciranje, distribuiranje, dajanje v najem, javna priobčitev, dajanje na voljo javnosti (internet), predelava ali vsaka druga uporaba tega avtorskega dela ali njegovih delov v kakršnemkoli obsegu ali postopku, vključno s fotokopiranjem, tiskanjem ali shranitvijo v elektronski obliki. Odstranitev tega podatka je kazniva.

Izdala in založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani,

Oddelek za prevajalstvo

Za založbo: Roman Kuhar, dekan Filozofske fakultete

Ljubljana, 2017

Prva izdaja, elektronska izdaja

Oblikovna zasnova: Kofein, d. o. o.

Prelom: Jure Preglau

Publikacija je brezplačna.

Publikacija je dostopna na: <https://e-knjige.ff.uni-lj.si>

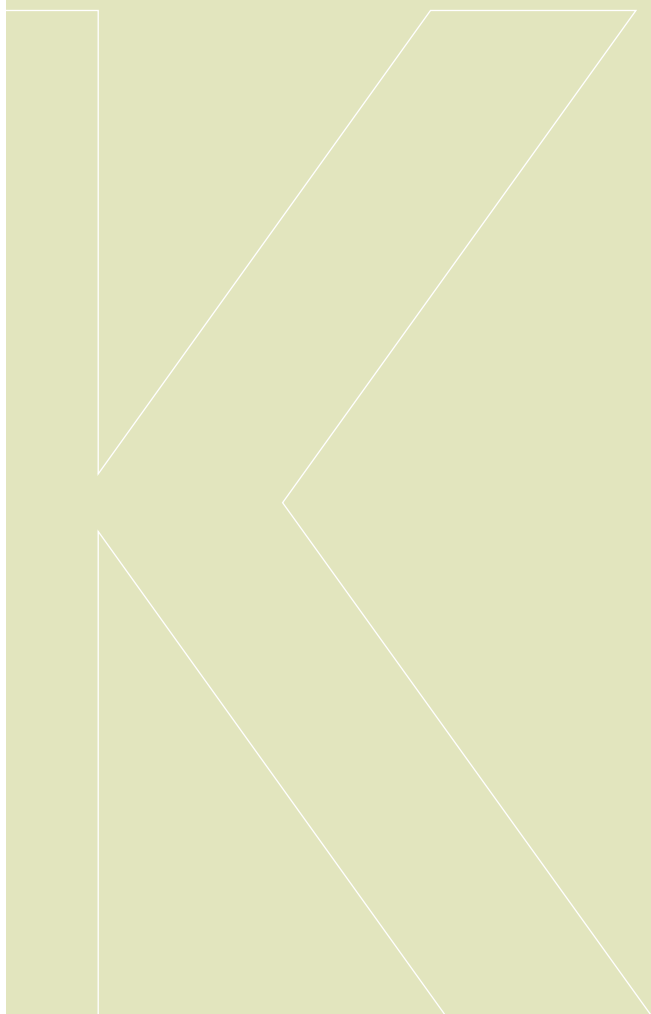
DOI: 10.4312/9789612379889

Kataložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=292908544

ISBN 978-961-237-987-2 (epub)

ISBN 978-961-237-988-9 (pdf)

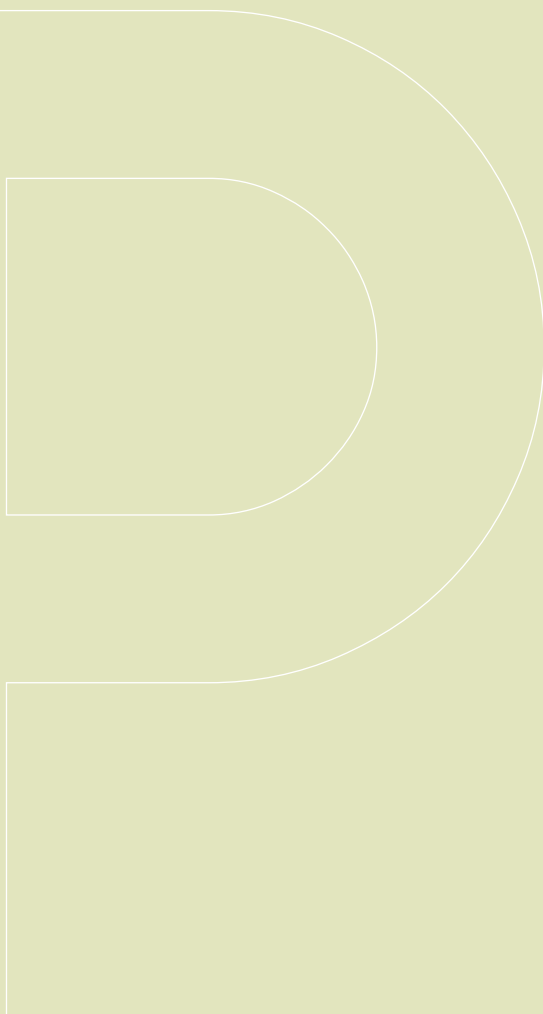
Kazalo vsebine



	Predgovor	8
1	Uvod	12
1.1	Splošno in strokovno v jeziku	13
1.2	Razmerje med strokovnim in splošnim jezikom	14
1.3	Predmet terminološke vede	18
1.4	Temeljne razlike med terminologijo in leksikologijo	19
2	Pojmi in poimenovanja	22
2.1	Pojmovno načelo	23
2.2	Razmerja med pojmi	27
2.2.1	Hierarhična razmerja	28
2.2.2	Strokovno specifična razmerja	30
2.3	Pojmi v medjezikovnih stikih	31
2.4	Terminologija in semantika: od slovarja do ontologije	32
3	Prepoznavanje in opis terminološkosti	36
3.1	Klasične in pragmatske definicije termina	37
3.2	Besedilni pristopi k terminološkosti	37
3.3	Oblika terminov	39
3.4	Terminološke variacije	42
3.5	Terminološkost s stališča uporabnikov	46
4	Terminotvorje	50
4.1	Prevzemanje	51
4.1.1	Izlastnoimensko izrazje	53
4.1.2	Kalkiranje	54
4.2	Tvorjenje novih poimenovanj v okviru lastnega jezika	55
4.3	Prevajanje kot terminotvorna dejavnost	56
5	Terminografija	58
5.1	Terminografske dejavnosti na Slovenskem	58
5.2	Struktura terminografskega vnosa	62
5.3	Definicije in razlage pojmov	65
5.4	Terminografija za prevajalske namene	67
6	Standardizacija terminoloških virov	70
6.1	Mednarodni terminološki standardi	71
6.2	Standardizacija v Sloveniji	75

7	Računalniško podprta terminografija	78
7.1	Gradnja specializiranih korpusov	78
7.1.1	Vrste korpusov	78
7.1.2	Reprezentativnost	82
7.1.3	Homogenost	85
7.1.4	WebBootCaT	86
7.2	Obdelava korpusa	88
7.2.1	Označevanje	88
7.2.2	Obdelava korpusa z orodjem Wordsmith Tools	90
7.2.2.1	Besedni sezname	91
7.2.2.2	Konkordance	96
7.2.2.3	Ključne besede	97
7.3	Samodejno luščenje terminologije	99
7.3.1	Statistični pristopi k iskanju terminov	101
7.3.1.1	TF-IDF	102
7.3.1.2	Drugi indikatorji terminološkosti	103
7.3.1.3	Statistike za luščenje kolokacij	104
7.3.2	Luščenje terminov s pomočjo oblikoskladenjskih vzorcev	109
7.3.3	Iskanje prevodnih ustreznice	110
8	Ustvarjanje terminoloških baz	112
8.1	Ustvarjanje terminoloških baz s programom SDL Trados Multiterm	113
8.2	Pogoste napake	120
	Stvarno kazalo	122
	Viri	126

Predgovor



S strokovnim izrazjem se občasno, pogosto ali vsakodnevno srečuje vsak uporabnik jezika. Prav vsaka človeška dejavnost je povezana z rabo specializiranih izrazov, ki se po svoji obliki sicer morda ne razlikujejo od običajnih besed, a so zaradi umeščenosti na določeno strokovno področje rabljeni terminološko. Namen pričujoče knjige je predstaviti teoretična izhodišča za profesionalno – ali tudi ljubiteljsko – ukvarjanje s terminologijo in nato iz njih izpeljati metodologijo za računalniško podprto terminografijo.

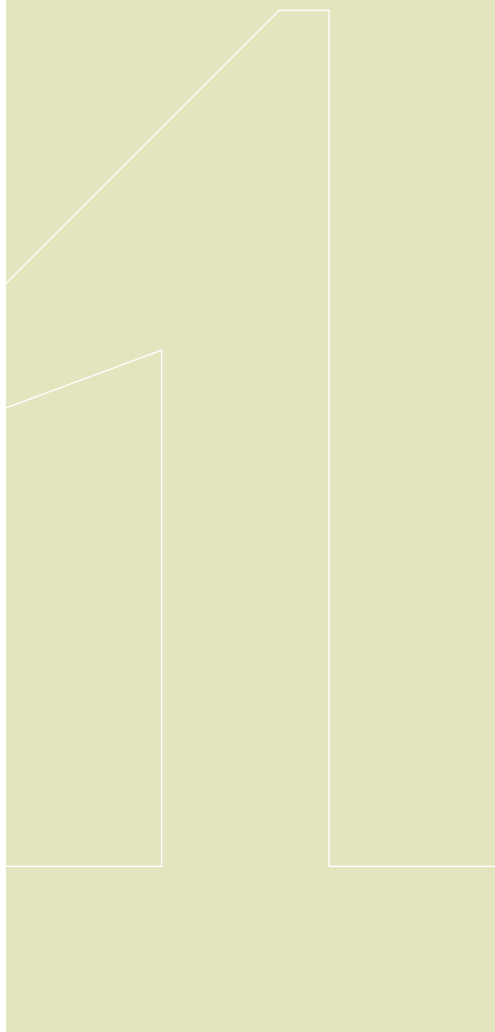
S terminologijo se ukvarjajo različne skupine uporabnikov, na eni strani gotovo strokovnjaki v okviru sporočanja specializiranih vsebin, prevajalci in tolmači ob njihovem prevajanju, terminografi ob gradnji terminoloških slovarjev in baz, lektorji in uredniki v različnih fazah revizije besedila, dokumentalisti in bibliotekarji ob klasifikaciji gradiva, računalničarji ob razvijanju sistemov za poizvedovanje po besedilnih zbirkah in tehnologij znanja.

Prav vsem omenjenim skupinam uporabnikov je skupno, da se prek terminologije uporablja, ureja, beleži, prenaša in ustvarja znanje. Povezava med strokovnim izrazjem in pojmovnimi sistemi človekovega vedenja je tako neposredna in primarna, da so se s terminologijo nekoč ukvarjali izključno znanstveniki pretežno naravoslovnih ved, v 30. letih 20. stoletja pa so se jim pridružili še predstavniki tehniških ved, vsekakor pa je povezava terminologije z jezikoslovjem in v okviru tega predvsem z leksikologijo razmeroma sodoben pojav. Iz te navidezne sorodnosti – saj se tako terminologija kot leksikologija v določeni stopnji ukvarjata z besedami – izhaja tudi največ konceptualnih in metodoloških zmot, ki jih skuša pričujoča knjiga vsaj v določenem segmentu pojasniti in morda vnaprej preprečevati. Pri tem je posebna pozornost namenjena večjezikovnemu vidiku, s katerega terminologijo obravnavajo prevajalci in terminografi večjezikovnih terminoloških zbirk za prevajalske namene.

V zadnjem delu, kjer so predstavljene nekatere korpusne in računalniške metode terminološkega dela, se je bilo treba že vnaprej sprijazniti z dejstvom, ki je vse bolj očitno tudi na področju klasične terminografije, in sicer da so knjige v tiskani obliki obsojene na zastarelost že v trenutku svojega izida. Programska orodja, spletne strani in računalniške tehnologije se navadno povsem zamenjajo vsakih nekaj let, a vsaj osnovna načela bodo – upajmo – uporabniku koristna. Predvsem na področju korpusne terminografije pa se je v zadnjih letih zgodil preboj v tem smislu, da v različnih okoljih nastajajo specializirani korpusi in na njih temelječe terminografske zbirke, strokovna javnost pa korpusnih metod ne obravnava več kot novotarijo. Pričujoča knjiga želi tako s teoretičnimi izhodišči in praktičnimi napotki spodbuditi razvoj predvsem v tej smeri.

Iskreno se zahvaljujem vsem, ki so z idejami, pogovori, popravki, vprašanji ali preprosto s spodbudnim vzdušjem pripomogli k nastanku te knjige: Vojku Gorjancu za dolgoletno kolegialno mentorstvo, prvo branje in recenzijo, študentom Oddelka za prevajalstvo FF za diskusije med predavanji in sicer, kolegom s taistega oddelka za plodno delovno vzdušje, Apoloniji Gantar za recenzentske pohvale, še najtopleje pa seveda svoji ljubi družini.

1 Uvod



Pričujoča knjiga se ukvarja s terminologijo, se pravi s tistimi izrazi v besedilu, s katerimi označujemo specializirane pojme. Preden pa se nadrobneje posvetimo vprašanju, kaj je termin in kako ga spoznamo, skušamo na kratko predstaviti okolje, v katerem utegnemo srečati specializirane izraze, se pravi strokovni diskurz oziroma strokovno besedilo.

1.1 SPLOŠNO IN STROKOVNO V JEZIKU

Jezikovna komunikacija opravlja številne sporočevalne naloge in največkrat gre pri tem za sporočanje o določeni temi ali področju. Takorekoč vsaka tema v komunikaciji pomeni, da bomo pri govorjenju ali pisanju o njej uporabili izraze, povezane s to temo. Če je tema ozko specializirana, bomo v komunikaciji uporabili specializirano izrazje, ki mu rečemo tudi terminologija, hkrati pa se bomo v takšen, se pravi specializiran, diskurz podali le ob domnevi, da sogovornik ali bralec uporabljeno izrazje pozna.

Specializirani oziroma strokovni diskurz se od *splošnega* torej najbolj opazno razlikuje po izrazju, čeprav imajo – kot bomo nadrobneje videli kasneje – nekateri strokovni jeziki tudi druge lastnosti, po katerih odstopajo od t.i. splošnega. Kriteriji, po katerih intuitivno ločimo strokovni jezik od splošnega, so denimo tile:

Komunikativna situacija

Te se ločijo glede na raven predznanja, ki ga izkazujejo ali predpostavljajo udeleženci v diskurzu. Komunikacija je tem bolj specializirana, čim več strokovnega znanja o tematiki delijo udeleženci.

Razumljivost

Če smo kot laik udeleženi v strokovnem diskurzu, nam raba specializiranega izrazja otežuje razumevanje. Tudi v prevajalskih situacijah raven strokovnosti besedila pogosto ocenjujemo po njegovi (ne)razumljivosti.

Besedilna vrsta

Pri nekaterih besedilnih vrstah intuitivno pričakujemo več specializiranega izrazja kot pri drugih. V znanstvenem članku bomo navadno našli več specializiranih terminov kot v novinarskem prispevku o isti temi, objavljenem v dnevnem časopisu.

A čeprav se nam v določenih situacijah, še posebej takih, kjer raven specializiranosti diskurza presega naše pričakovanje in/ali predznanje, zdi, da je prepoznavanje terminov in njihovo razlikovanje od običajnih besed enostavno in nedvoumno, se

pri številnih dejavnostih, ki se ciljno ukvarjajo s terminologijo, to izkaže za težko nalogo. Termini so namreč pretežno besede in besedne zveze, zato je ukvarjanje z oblikoslovnimi ali etimološkimi lastnostmi terminov v resnici ukvarjanje s temi lastnostmi besed. Če problem obrnemo in se vprašamo po besedah s specializiranim ali terminološkim pomenom, smo sicer že bližje, a kot bomo podrobneje spoznali v 3. poglavju, tudi pragmajezikoslovnim ali besediloslovno usmerjenim definicijam termina ne uspeva postaviti trdnih meja med enimi in drugimi.

1.2 RAZMERJE MED STROKOVNIM IN SPLOŠNIM JEZIKOM

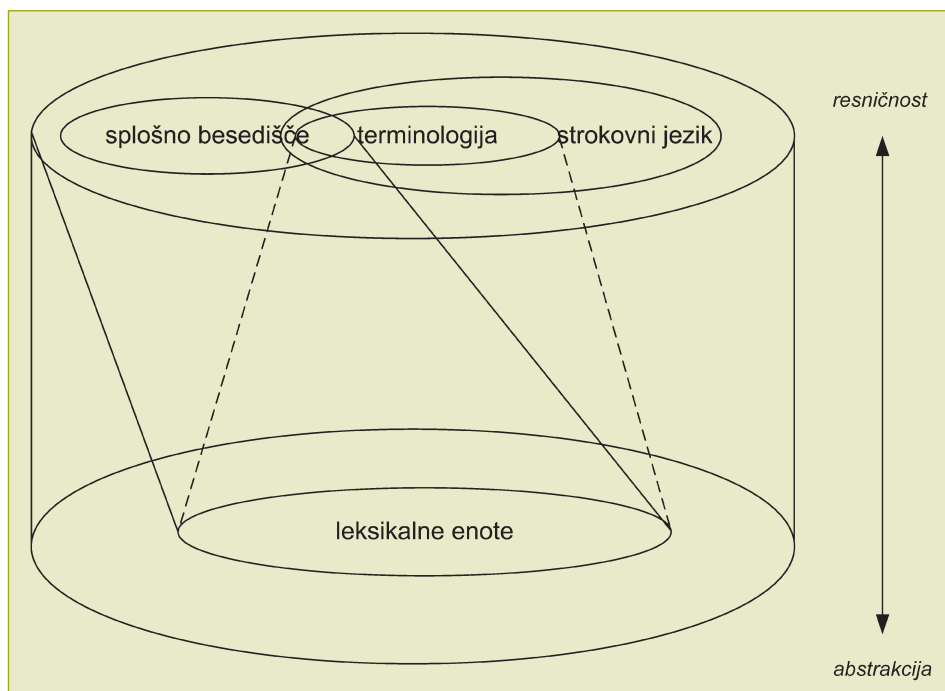
Čeprav imamo na voljo precej intuitivnih kriterijev, ki nam pomagajo razlikovati med strokovnim in splošnim jezikom, je ta dva pojma težko nedvoumno opredeliti. V slovenščini nam dela težave že izraz strokovni jezik, ki ga razumemo kot jezik stroke, se pravi jezik neke uveljavljene, posebno znanje zahtevajoče človeške dejavnosti.¹ Po tej razmeroma široki definiciji stroke bi mednje sicer lahko šteli tudi športne in druge prostočasne dejavnosti, še vedno pa bi težko našli ljudi, ki bi govorili o stroki balinanja ali likanja.

Mnogh človeških dejavnosti torej ne moremo uvrščati med stroke, pa vendar tudi te dejavnosti uporabljajo svojo enoznačno terminologijo ali svoj žargon. V angleščini za strokovni jezik ni pravega prevoda, v rabi pa sta izraza *sublanguage* in *special language*; oba brez sporne navezave na stroko in bolj neobremenjena z uveljavljenostjo področja, na katerem se uporabljata. Ker bi v slovenščini izraz *podjezik* utegnil biti narobe razumljen kot slabšalen, bi bilo morda najprimerneje govoriti o posebnem, specialnem ali področnem jeziku. Zaradi uveljavljenosti in domačnosti pa v okviru te knjige še naprej uporabljamo strokovni jezik, pri katerem imamo v mislih vse oblike specializiranega izražanja, tipičnega za stroke, znanosti, vede, področja in dejavnosti, ki vključujejo posebno znanje in kjer je mogoče razlikovati med laiki in strokovnjaki.

Kageura (2002: 13) predstavlja strokovni jezik kot del jezika, terminologijo pa kot del strokovnega jezika. Poleg terminologije sestavlja splošni jezik tudi drugo, splošno besedišče, ki se s strokovno terminologijo deloma pokriva. Tako splošno kot strokovno besedišče pa sestavljata nabor leksikalnih enot jezika (Slika 1). Splošni in strokovni jezik se po tej shemi nahajata na ravni resničnosti, saj sta udejanjena v diskurzu. Leksikalne enote so na drugi strani po svoji esenci abstraktne in šele raba jih umesti v splošni ali strokovni diskurz.

¹ Po SSKJ celo le *gospodarska dejavnost* ali *znanstvena panoga*.

Slika 1: Položaj terminologije znotraj jezika (Kageura 2002: 13)



Cabr  (1998: 58) pravi, da je jezik sestavljen iz številnih podkodov, ki jih govorci uporabljajo v skladu s svojimi izraznimi potrebami in z naravo komunikacijske situacije. Kljub tej raznovrstnosti pa v vsakem jeziku obstaja skupni nabor enot in pravil, ki jih poznajo vsi govorci. Ta nabor naj bi sestavljal splošni jezik, enote splošnega jezika pa se uporabljajo pretežno v nezaznamovanih situacijah.  e v naslednjem odstavku pa Cabr  ugotavlja, da takšna opredelitev splošnega jezika zanemara dejstvo, da govorci uporabljajo splošni jezik v druga nih situacijah kot strokovni jezik.

Arntz in Picht (1995: 20) poudarjata temeljno razliko med splošnim in strokovnim jezikom: strokovni jezik kljub vsem svojim posebnostim ne more obstajati brez splošnega jezika, saj uporablja njegov slovni ni in besedni inventar, medtem ko splošni jezik brez dvoma lahko obstaja tudi sam zase. Pri zgornjem širokem pojmovanju strokovnega jezika pa je tudi samostojnost – ali sam obstoj – splošnega jezika na majavih nogah, saj, kot piše Klinar (2004: 103), »vse  loveške dejavnosti nazadnje porabijo vse besediš e jezika, [in] smo v skušnjavi, da bomo vse besediš e – razen funkcionalnih besed (ali v jezikoslovju nemara celo te?) – imeli za tehni ne termine.«

Tudi če se tej skušnjavi upremo, in sicer ponavadi tako, da s terminološkega vidika obravnavamo le eno področje naenkrat, je dobro poznati še druge lastnosti strokovnega jezika, ki ga razlikujejo od splošnega in ki ne zajemajo le strokovnega besedišča. Cabré (1998: 65) navaja naslednje kriterije:

1. O strokovnem področju govorimo, kadar to ni del govorčevega splošnega znanja ampak znanja, pridobljenega v okviru specifičnega učnega procesa.
2. Govorec, ki ima takšno znanje, je strokovnjak in uporablja strokovni jezik. Pri strokovni komunikaciji razlikujemo med izvornimi in ciljnim uporabniki strokovnega jezika, kjer mora izvorni uporabnik dejavno obvladati strokovno področje, ciljni uporabnik pa je lahko tudi poučeni nestrokovnjak, ki pasivno sprejema strokovno komunikacijo v procesu usvajanja znanja.
3. Strokovna komunikacija je navadno formalna in se najpogosteje pojavlja v poklicnih, strokovnih ali znanstvenih situacijah.
4. Strokovni jezik ima številne jezikovne (enote in pravila) in besedilne značilnosti (besedilne vrste).
5. Strokovni jezik nikakor ni jasno zamejena podmnožica, temveč dopušča številne različice glede na rabo in komunikacijsko situacijo. Spremenljivi parametri pri tem so:
 - a. stopnja abstrakcije, ki je odvisna od strokovnega področja, sprejemnikov informacij in govorčevega namena,
 - b. komunikativni cilj, ki določa različico besedilne vrste,
 - c. geografski, zgodovinski ali socialni dialekt,
 - d. osebni stil.
6. Strokovne jezike povezujejo mnogi pragmatični in jezikovni dejavniki, zato je mogoče o njih govoriti tudi z notnega stališča kot o podmnožici splošnega jezika.
7. Strokovni jeziki so podsestavi v okviru jezika kot takega. Pri tem se delno prekrivajo s splošnim jezikom, in sicer ne le tako, da prevzemajo večino njegovih pravil, ampak med njima neprestano teče obojestranska izmenjava jezikovnih enot in konvencij.

Pri nekaterih besedilih smo morda v dvomih, ali gre za strokovno besedilo ali ne. Vzemimo za primer spodnji izsek iz besedila *Živinozdravstvo*.² Gre brez dvoma za besedilo, ki je namenjeno kmetom in živinorejcem v najširšem smislu, pri čemer je iz besedila razvidna časovna – in s tem tudi kulturna – oddaljenost od današnjega časa. Kljub tej oddaljenosti pa o strokovnosti besedila ne more biti dvoma in če preverimo zgornje kriterije, jih hitro najdemo nekaj, ki ustrezajo.

² Rohlwes, Johann Nikolaus (1856) *Domače živinozdravstvo v boleznih konj, govedjā, ovac, prešičev, koz in psov, ali nauk, kako mora kmetovavec svojo živino rediti, ji streči, jo kermi in ozdravljati* [prevedel Karl Robida]. V Celovcu: Janez Leon.

Čeprav je živinoreja razširjena dejavnost med kmeti, z njo in z zdravjem živali povezana znanja niso prirojena, ampak si jih je treba pridobiti z ustreznim učnim procesom. Rejci konj nedvomno uporabljajo strokovno terminologijo, v spodnjem besedilu denimo *celak*, *bencelj*, *krona*. Raven specializiranosti je prilagojena tipu diskurza, se pravi komunikaciji strokovnjak – polstrokovnjak ali laik. Besedilo nadalje spoznamo za strokovno zaradi čustvene nezaznamovanosti, predvsem pa tudi zaradi načina, kako avtor uvaja in razlaga strokovne izraze: *pišalo* – *podkolenska cev*, *krona* – *žilnati svitek*, *dila* – *spodnja plat kopita*.

PERVI RAZDELEK.

Od reje, strežbe in kermenja konj.

Konjoreja je poglavitno sredstvo, blagostanje kmetovavca podpirati in ohraniti, ker si tako za domačo rabo potrebne in dobre konje brez velikih stroškov priredi, in verh tega lepe denarje prejme, ako zamore kakega konja prodati. Torej bo gotovo vso pridnost na konjorejo obrnil, in po tem, kar se mu tukaj nasvetuje, se ravnal.

Narprej mora kmet za dobre, terdne in izverstne kobile skrbeti, in verh tega paziti, da se k njim celaki ali žrebci enake vlastitosti pripuščajo. Lastnosti dobrega celaka, na ktere vsak trgovavec s konji gleda, so pa sledeče:

Dobrega celaka glava je majhena in celem životu primerna; ušesa sta majhena in ne daleč saksebe; očesa velika, zvišana in bez madeža; vrat je dolg, pod grivo ne preširok in ne po mulno raztegnjen. Persi morajo biti široke, ravno tako morajo zgornji deli sprednjih nog debeli in krepki biti; ako ste sprednji nogi tanki in šibki, nima konj nobene moči. Koleno mora biti od spredej široko, *pišalo* ali *podkolenska cev* pa in njena kita se mora od strani široka viditi. *Bencelj*, ali tisti del noge, ki je med *benceljnovim* členom in *žilnatim* svitkom ali *krono*, mora prav kratek biti, in ako je daljši, se per stopanju ne sme kriviti, kar bi dokazovalo slabost tega dela. Rožena stena mora biti krog in krog gladka brez obročev in brez podolgastih poklin. Spodnja plat kopita ali *dila*

1.3 PREDMET TERMINOLOŠKE VEDE

Terminološka veda se ukvarja s *termini* kot poimenovanji za specializirane pojme določenega strokovnega področja. *Pojmi* niso le predmetnosti določenega področja, ampak tudi procesi, lastnosti, dejavnosti ali načini. To poudarjamo predvsem zato, ker za termine z oblikoslovnega stališča navadno pričakujemo, da so samostalniki in samostalniške zveze, kar pa za mnoga strokovna področja ne drži, saj med njimi najdemo tudi glagole, pridevnike, prislove. Inventar jezikovnih poimenovanj pojmov neke stroke imenujemo tudi *terminologija*, na primer geološka, medicinska, planinska terminologija.

Obenem se tudi terminološka veda imenuje terminologija. Na to dvoumnost opozarja standard ISO 1087:

terminology: Set of terms representing the system of concepts of a particular subject field.

terminology science: The scientific study of the concepts and terms found in special languages.

terminography: The recording, processing and presentation of terminological data acquired by terminological research.
(ISO 1087)

Čeprav se terminološka veda ukvarja predvsem s specializiranim izrazjem, je pri dejanskem terminološkem delu – pa naj gre za slovaropisne, prevajalske ali kake druge namene – izrazje nemogoče obravnavati ločeno od drugih ravni, denimo skladišne, besediloslovne in pragmajezikoslovne. O tem, kaj v določenem besedilu razumemo kot termin, nam namreč največ povedo značilnosti besedila in okoliščine njegovega nastanka, seveda pa tudi neposredno besedilno okolje izraza.

Naloge terminološke vede zajemajo (Felber 1984, Arntz in Picht 1995):

- preučevanje, zbiranje in obdelavo terminologij posameznih strok,
- preučevanje pojmov, pojmovnih sklopov in razmerij ter njihovo poimenovanje z ustreznimi termini,
- preučevanje in usklajevanje poimenovanj v večjezikovnem kontekstu,
- pripravo in izdajanje terminoloških priročnikov.

Klasična terminologija je po definiciji tudi normativna, se pravi ne le opazuje in opisuje, ampak tudi predpisuje terminološko rabo. Svojo predpisovalno vlogo izpolnjuje na različne načine:

- svetuje prednostne in neprednostne izraze in udejanja načelo enoznačnosti,

- z izdajanjem normativnih priročnikov skrbi za ustaljenost in doslednost izrazja,
- sproža in nadzoruje postopke terminološke standardizacije.

Danes imamo za večino strokovnih področij na voljo tudi pisna gradiva v elektronski obliki, ki omogočajo lažje in hitrejše zbiranje terminoloških podatkov. Sodoben pogled na terminološko in predvsem terminografsko dejavnost tako poudarja pomen korpusnega pristopa, ki se je sicer že pred časom uveljavil tudi v splošni leksikografiji. Pri takem pristopu je normativno delovanje mogoče le na posreden način, čeprav je z vidika uporabnosti terminoloških priročnikov še vedno potrebno. Kadar za isti pojem obstaja več izrazov, je primerno in zaželeno, da terminograf poda priporočila glede prednostne rabe. Korpusna terminografija je usmerjena v opazovanje in opisovanje jezikovnih pojavov na podlagi kvantitativnih dokazov, zbranih iz velikih besedilnih zbirk, ki služijo kot vzorec jezika ali strokovnega jezika. Tako korpusna terminologija lahko pomaga ugotavljati, kakšna je dejanska jezikovna raba na strokovnem področju in za to ponuja tudi statistične dokaze, a za udejanjanje predpisovalne vloge je te vsekakor potrebno še kvalitativno ovrednotiti.

S tem pa korpusna terminologija premošča številne vrzeli, ki jih klasični terminološki vedi danes lahko očitamo. Predvsem tu mislimo na nezmožnost terminološke vede, da bi v naglo spreminjajočem se svetu globalnega tehnološkega razvoja še izpolnjevala zgoraj navedene naloge, še posebej pa na njeno neodzivnost na globoke spremembe v načinih in medijih strokovne komunikacije. Te namreč korenito posegajo tudi v naloge terminološke vede, ki – v zgoraj navedeni obliki – ne morejo več zadovoljevati potreb uporabnikov terminologije.

1.4 TEMELJNE RAZLIKE MED TERMINOLOGIJO IN LEKSIKOLOGIJO

Kot se terminologija ukvarja s termini in terminotvorjem, se tudi leksikologija ukvarja z leksemi in postopki tvorjenja leksikalnih enot. Obema vedama je skupno, da se – prek terminov in leksemov – ukvarjata z besedami, saj so tako termini kot leksemi sestavljeni iz nizov ene ali več besed, kljub temu pa je med njima nekaj bistvenih razlik, ki nam pomagajo tudi bolje opredeliti namen terminološke vede. Najprej je tu razlika med predmetom preučevanja pri prvi in pri drugi, se pravi razlika med terminom in leksemom.

Termin je, po zgornji opredelitvi, poimenovanje strokovnega pojma, in kot tak nastopa v strokovni komunikaciji, leksem pa je besedna enota jezika kot celote in se prav tako lahko pojavlja v splošni, nespecializirani komunikaciji. Ker si v strokov-

nem diskurzu želimo čim manj nejasnosti, je za termine zaželeno, da so enoznačni oziroma enopomenski, vsaj znotraj ene stroke. Pomen termina mora biti mogoče jasno opredeliti s strokovno definicijo. Nasprotno so leksemi pogosto večpomenski; na splošno je večpomenskost zanimiv jezikovni pojav, brez katerega si ne moremo zamisliti besednih iger in številnih stilističnih figur. Pomen leksikalnih enot je okvirno opredeljen v splošnem razlagalnem slovarju, pri čemer se ta glede na sobesedilne okoliščine lahko tudi precej spreminja.

Nadalje so termini včasih prepoznavni že po svoji obliki in zgradbi, saj vključujejo več tujejezičnega gradiva in lahko vsebujejo tudi numerične znake, simbole in druge neabecedne sestavine. Poleg tega naj bi bila večbesedna terminološka enota ustaljena in čimbolj nespremenljiva, medtem ko je pri večbesednih leksikalnih enotah uporabniku prepuščeno več svobode. Da so tudi terminološke enote spremenljive in se – pogosto izključno iz slogovnih razlogov – pojavljajo v več variacijah, bomo podrobneje opisali v nadaljevanju.

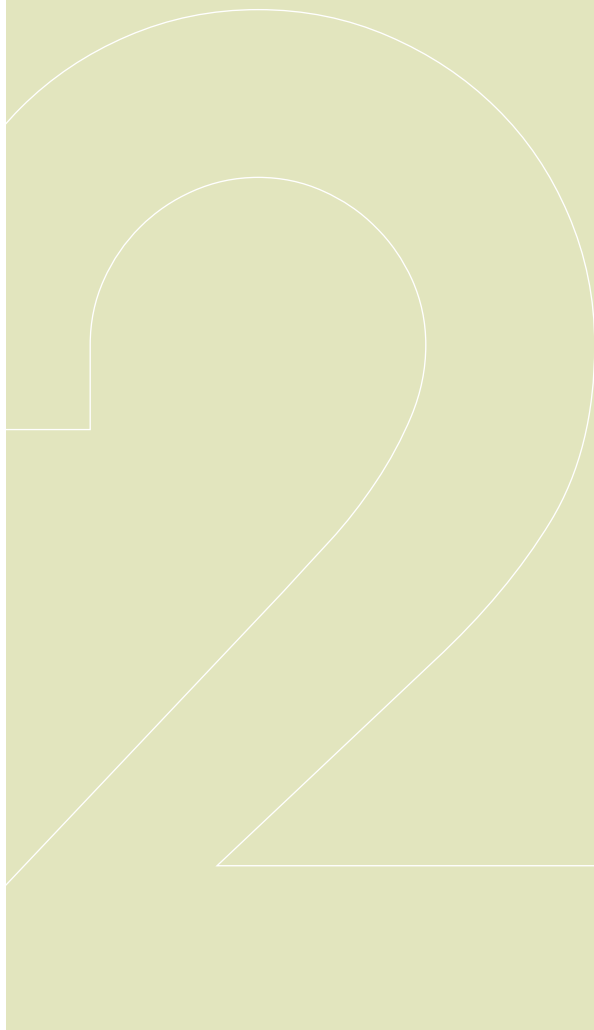
Glavne razlike med terminološkimi in leksikalnimi enotami povzema Tabela 1.

Tabela 1: Razlike med terminološkimi in leksikalnimi enotami

terminološka enota	leksikalna enota
• ena od možnih reprezentacij pojma • en pomen, v besedilu in v stroki • pomensko določena z definicijo • lahko vsebuje tudi numerične znake, simbole, logotipe itd. • se loči od neterminoloških enot po specializirani referenčnosti	• več možnih pomenov (konotativni vs. denotativni) • ekspresivna raba • pomensko opisana v splošnem slovarju; pomensko polje se spreminja v odvisnosti od sobesedila • oblikovno svobodnejša • načeloma vsebuje zgolj besedni material

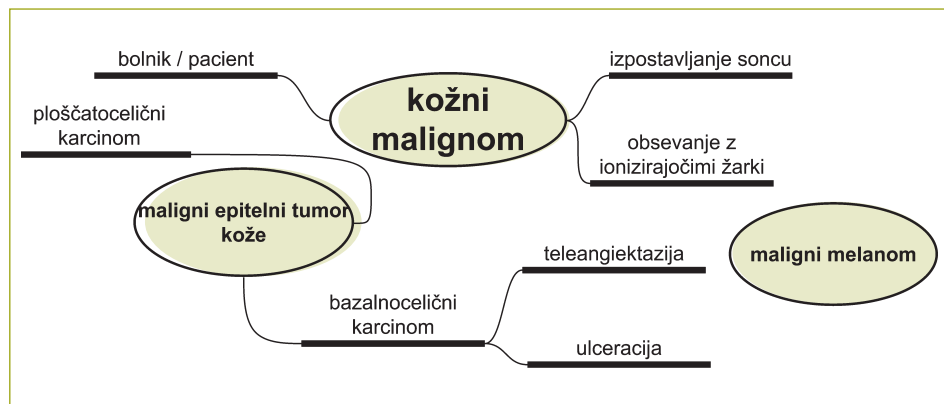
Zdaj ko smo primerno poudarili razliko med terminom in leksemom, moramo prej povedano postaviti v realnejšo luč. Kot smo namreč deloma razložili že v uvodnem poglavju, formalne in nedvoumne ločnice med terminom in leksemom ni oziroma je ta izključno funkcionalna. Termini so večinoma leksemi, zato se pri obeh vedah slej ko prej ukvarjamo z besedami in besednimi zvezami. Da je srž terminologije predvsem dostopanje do znanja, nadrobneje pojasnjuje naslednje poglavje.

2 Pojmi in poimenovanja



Pojmi predstavljajo najmanjše elemente, nekakšne atome, iz katerih je zgrajeno specializirano znanje. Pojma nikdar ne moremo obravnavati povsem ločeno od širšega pojmovnega polja, ki mu pripada, saj so pojmi prek številnih in raznolikih *pomenskih razmerij* povezani v večrazsežnostno pojmovno mrežo. Za ponazarjanje kompleksnih pomenskih razmerij je naravni jezik pogosto precej omejeno orodje, zato se na številnih strokovnih področjih pojmovna polja na nejezikovni način predstavljajo s pomočjo slik, grafov, simbolov, shem ipd. (Galinski in Picht 1995).

Slika 2: Pojmovno polje okrog *kožnega malignoma*



Zgornji primer pojmovnega polja, ki se konstituira okrog pojma *kožni malignom* na podlagi znanstvenega besedila (Bartenjev in dr. 2006), izkazuje različna pojmovna razmerja, pogojena z značilnostmi in vzročno-posledičnimi povezavami na tem strokovnem področju. Tako ena veja sheme predstavlja hiponime ali podrejene pojme, saj maligni epitelni tumor kože in maligni melanom predstavljata vrsto kožnega malignoma, prvi pa se deli naprej na ploščatocelične in bazalnocelične karcinome. Druga razmerja na zgornji sliki pa niso hierarhična v smislu pod- ali nadpomenskosti, ampak tematska, na primer vzročno (izpostavljanje soncu *povzroča* kožni malignom), terapevtsko (obsevanje z ionizirajočimi žarki *je oblika zdravljenja* kožnega malignoma), simptomatsko (ulceracija *je eden od simptomov* pri rasti bazalnoceličnega karcinoma) ali subjektivno (bolnik *zboli za* kožnim malignomom).

2.1 POJMOVNO NAČELO

Kot rečeno je pojem miselna enota, zato je njegovo jezikovno poimenovanje ali termin le neke vrste nalepka, ki mu jo priredimo. Za isti pojem se lahko uporablja več poimenovanj v enem jeziku, v različnih jezikih pa je zanj lahko v uporabi na stotine izrazov. Pogosta je tudi situacija, da v nekem jeziku ni poimenovanja za določeni pojem, čeprav je ta jasno opredeljen in ga je mogoče definirati s sorodnimi pojmi. Če podamo primer, se na področju vojaštva v angleškem jeziku uporablja glagol *dock* za pristajanje plovila in glagol *land* za pristajanje zračnega plovila. Oba pojma, se pravi miselni predstavi, obstajata tudi v slovenščini, vendar razlikovanja pri poimenovanju ni; za oba pojma se uporablja izraz *pristajati*. Poleg tega se nekateri pojmi označujejo tudi s kratico, simbolom ali (kemijsko) formulo, pri nekaterih pojmi, ki označujejo denimo podjetja ali politične enote, pa so najbolj razpoznavni logotip, grb ali zastava.

Pojmovno načelo v terminologiji pomeni predvsem, da termine obravnavamo kot jezikovna poimenovanja pojmov in zatorej izhajamo iz pojmov kot osnovnih enot, ne pa iz leksikalnih lastnosti terminov. Za lekseme je namreč značilno, da imajo lahko več pomenov, v terminologiji pa pomen enačimo s pojmom. En pojem ima kot abstraktna miselna enota lahko torej le en pomen, ki pa mu seveda lahko pripišemo več poimenovanj. Prav tako ni nujno, da določeni termin označuje le en pojem. Za terminografsko prakso to pomeni, da so terminološke zbirke sestavljene iz vnosov, kjer en vnos ustreza enemu pojmu, pod isti vnos pa nato beležimo vsa možna poimenovanja tega pojma ter tudi vse ostale podatke, ki se nanašajo na pojem. Podrobneje strukturo terminoloških zbirk opisujemo v poglavju 5.

Slika 3 je zaslonski posnetek pojmovnega vnosa v Evrotermu, kjer sta za slovenski jezik navedeni dve konkurenčni poimenovanji za pojem *aneks/priloga*, raba pa je priporočena s kvalifikatorjem.

Slika 3: Primer iz Evroterma

EVROTERM	
večjezična terminološka zbirka	
CreationDate	16.06.1997
ChangeDate	16.06.2006
Subj	pravo
Subject	law JU
ID	67
SL	<u>aneks</u>
SL	<u>priloga</u>
UsageLabel	priporočeno
Reliability	4
Note	Z veliko začetnico pišemo, kadar je govor o konkretni prilogi v določenem sporazumu.
TermRef	Priročnik za projekte tesnega medinstitucionalnega sodelovanja
EN	<u>annex</u>
Definition	Matter complementary to the main text placed at the end of a document containing notes, statistical tables or other information.
TermRef	Reference Manual on Twinning Project

Slika 4 kaže primer iz ene najboljsežnejših terminoloških baz na svetu, pojmovnega sistema za področje medicine UMLS (Unified Medical Language System), ki združuje različne medicinske nomenklature in vire v prepleteno semantično omrežje. Pri danem primeru je iskani niz *HIV-1* privedel do pojma z oznako CUI (*Concept Unique Identifier*) C0019704, ki ima v angleščini 26 poimenovanj. Podrobnejši pregled seznama terminov nam razkrije, da gre v resnici za šest različnih terminov s številnimi variacijami v zapisu. Primer ponazarja pojmovno načelo v praksi, kjer je osnovni element pojem (*concept*), ki ustreza enemu pomenu.

Slika 4: Poimenovanja za pojem HIV iz medicinskega metatezavra UMLS

Concept: HIV-1
CUI: C0019704
Semantic Type: <u>Virus</u>
Synonyms:
<u>HIV-1</u>
<u>AIDS virus</u>
<u>HIV 01</u>
<u>HIV1</u>
<u>HIV1 - Human immunodeficiency virus type 1</u>
<u>HIV1 - Hum immunodef vir typ 1</u>
<u>HIV-I</u>
<u>HTLV-3 - Human T-cell lymphotropic virus type 3</u>
<u>HTLV-3 - Human T lymph virus 3</u>
<u>HUMAN IMMUNODEFICIENCY VIRUS 001</u>
<u>Human immunodeficiency virus 1</u>
<u>human immunodeficiency virus-1 HIV-1</u>
<u>human immunodeficiency virus HIV-1</u>
<u>Human immunodeficiency virus type 1 [Ambiguous]</u>
<u>human immunodeficiency virus type-1 HIV-1</u>
<u>Human immunodeficiency virus type I</u>
<u>Human immunodeficiency virus type I (organism)</u>
<u>human immunodeficiency virus type I HIV-1</u>
<u>Human immunodeficiency virus type I -RETIRED-</u>
<u>HUMAN IMMUNODEFIC VIRUS 1</u>
<u>HUMAN IMMUNODEFIC VIRUS TYPE 1</u>
<u>Human T-cell leukemia virus III</u>
<u>Human T-cell lymphotropic virus 3</u>
<u>Human T-cell lymphotrop virus3</u>
<u>human T lymphotropic virus III</u>
<u>Immunodeficiency Virus Type 1, Human</u>
Other languages:
Czech
<u>HIV-1</u>
Dutch
<u>Deficiëntievirus type 1, humaan immuno-</u>
<u>HIV 1</u>
<u>HIV-1</u>
<u>HIV-1 (humaan immunodeficiëntievirus, type 1)</u>
<u>HIV-LAV 1</u>
<u>Humaan immunodeficiëntievirus type 1</u>
<u>humane T-cel leukemie virus III</u>

Čeprav zgornji primer različna poimenovanja za virus HIV-1 niza kot sinonime, je pri pojmovnem načelu v navadi, da terminov, ki v enem jeziku označujejo isti pojem, ne imenujemo sinonimi. *Sinonimija* ali istoimenskost na ravni pojmov ne more obstajati, saj vsak pojem označuje svojo miselno enoto, na ravni terminov pa namesto o sinonimih govorimo o različnih poimenovanjih, ki jih skušamo na tak ali drugačen način tudi ovrednotiti s kvalifikatorji. Pojav več vzporednih poimenovanj za isti pojem je v terminologiji načeloma nezaželen, saj vodi do nejasnosti v strokovni komunikaciji. *Načelo enoznačnosti* v terminologiji tako narekuje, da si pri normativnem terminološkem delu prizadevamo priporočati le eno poimenovanje in ostala odsvetovati.

V praksi so vzporedna poimenovanja ali dvojnice, pa tudi terminološke variacije, neizogiben in nikakor ne nezaželen pojav. »Življenje« terminov je tesno povezano z evolucijo strokovnih področij, ki jim pripadajo (Jaquemin 2001:3), zato je normiranje terminologije v obliki klasičnih, tiskanih terminoloških zbirk vse manj smiselna dejavnost. Nove predmetnosti se razvijajo tako hitro, da bi nam – ob strogem spoštovanju načela enoznačnosti – že zdavnaj zmanjkalo poimenovanj, če termini ne bi prehajali med strokami in se prilagajali novim komunikacijskim funkcijam.

Za ameriško angleščino je takšna prehajanja statistično ovrednotil Losee (1995: 265-274), ki je znanstvene vede razdelil na »darovalke« in »prejemnice« glede na to, ali več izrazov, kot si jih same sposojajo, posojajo drugim področjem ali obratno. Njegova raziskava, ki je sicer temeljila na nekoliko skromnem besedilnem vzorcu, je potrdila izhodiščno hipotezo, da so t. i. eksaktne vede (*hard sciences*) pretežno darovalke, manj oprijemljive vede (*soft sciences*) pa prejemnice.

Če se vrnemo k pojmu, se za njegovo opredelitev, definicijo in umestitev v pojmovno polje uporabljajo njegove lastnosti, predvsem tiste, ki olajšujejo njegovo razmejitev od nadrejenega pojma, podrejenih in sorodnih pojmov. Lastnosti pojma lahko zajemajo materialne značilnosti, kot so barva, oblika, velikost in sestava (*krožna* (žaga), *enakostranični* (trikotnik)), značilnosti, ki se nanašajo na namen, denimo način uporabe, delovanje, učinek (*vzvratna* (prestava), *kreditna* (kartica)); izvor pojma, vključno z imenskimi sestavinami pri izumih, deželah izvora ali proizvajalcih (*Evklidska* geometrija, *Citroën* C6), in druge razlikovalne lastnosti (prim. Felber 1984: 117-123).

Iz zgornjih odstavkov se morda ponuja zaključek, da so pojmi nekakšni otočki v morju človeškega znanja, z jasno razvidnimi obalami, ki se med seboj sicer povezujejo, a so sami po sebi razmeroma stalni. V resnici se pojmi tako znotraj strok kot v miselnih svetovih uporabnikov neprestano spreminjajo. Pri vsakem poskusu

opredelitve ali zamejitve pojma se namreč srečamo z vprašanji vidika, namena in aplikacije, ki ključno vplivajo na naše pojmovanje sveta. Pojem *Sony Ericsson 800i* je lahko z vidika uporabnika telefon z zmogljivim fotoaparatom, z vidika serviserja model, ki se izjemno hitro poškoduje ob stiku z vodo in z vidika prodajalca Nokijinih telefonov nevarni konkurent. Na področju medicine je nešteto pojmov, ki so lahko z enega vidika simptomi, uporabljeni v diagnostičnem postopku, z drugega pa bolezenska stanja, ki zahtevajo terapevtske ukrepe. Tudi na individualni ravni nam vsako novo spoznanje nekoliko posodobi pojmovne sisteme, ki si jih gradimo v možganih.

Pojmovni sistemi in z njimi povezana (večjezična) poimenovanja nam danes služijo predvsem v okviru tehnologij znanja, kjer lahko sodobna terminološka veda ponuja svoj instrumentarij za učinkovitejšo gradnjo modelov znanja. Aplikacij, ki delujejo na temeljih formaliziranega znanja, je precej – ekspertni sistemi ali sistemi za podporo odločanju, odgovarjanje na vprašanja (*Question Answering*), še več pa si jih zamišljamo v prihodnosti. Tako se že nekaj let razvija t. i. semantični splet, ki naj bi informacijske tehnologije povzdignil na novo raven s pomočjo ontologij in drugih semantičnih virov. Čeprav se utegne medtem razvoj obrniti še v kako nesluteno smer, pa pojmovni sistemi in terminološko modeliranje svojo moč pokažejo šele v okviru določene (inteligentne) aplikacije.³

2.2 RAZMERJA MED POJMI

Večina znanstvenikov se strinja, da je znanje človekova stvaritev. S pomenskimi razmerji nove predmetnosti navezujemo na znanje, ki ga že imamo, in tako ustvarjamo novo znanje. Znanje se gradi s pomočjo smiselnega učenja (za razliko od učenja na pamet ali pogojnega učenja), to pa poteka takrat, kadar učeči namerno išče povezave med novimi informacijami in sorodnim že usvojenim znanjem (Ausubel et al., 1978). Pomenska razmerja so zato pomemben del terminoloških virov, ki klasični zbirki na tak ali drugačen način predstavljenih izrazov določenega področja dodajajo novo razsežnost v smeri baze znanja. Brez povezav med pojmi je terminološka baza zbirka samostojnih drobcev predvsem jezikoslovnih podatkov. Z vključevanjem pojmovnih razmerij začnejo ti drobci tvoriti pojmovne mreže in omrežja, z ustrezno formalizacijo vsebovanih podatkov pa se klasična terminološka zbirka tako približa večnamenskemu semantičnemu leksikonu ali ontologiji, primerni za uporabo v semantičnih aplikacijah.

³ To se med drugim kaže že skozi poimenovanje področja, saj se namesto o terminološki vedi (*Terminology Science*) v zadnjem desetletju raje govori o terminoloških tehnologijah (*Terminology Engineering*) oziroma upravljanju terminologije (*Terminology Management*).

2.2.1 Hierarhična razmerja

Pojmovna polja se ustvarjajo z razmerji med pojmi. Najenostavnejša razmerja so hierarhična ali taksonomska in obsegajo nad- in podpomenskost (hiper- in hiponimijo) ter razmerje del-celota/celota-del (mero- in holonimijo). Čeprav so hierarhična razmerja na nekaterih področjih temeljnega pomena za urejanje znanja, denimo na področju biologije za razvrščanje organizmov v taksonomijo živih bitij ali na področju tehnike za opisovanje naprav in njihovih sestavnih delov, pa se na številnih drugih strokovnih področjih izkaže, da je za urejanje pojmov v pojmovne sisteme nujno opredeliti tudi vidik oziroma kriterij, po katerem razvrščamo.

To je razvidno iz naslednjega primera tabelarične ureditve podpomenk pojma *zračno plovilo na motorni pogon* ali angl. *airplane* (Tabela 2). Tabela je povzeta po Felber (1986: 143).

Tabela 2: Podpomenke pojma zračno plovilo na motorni pogon, urejene po vidikih

a Vzletna in pristajalna površina	b Hitrostni razpon	c Pogon	d Dolet	e Uporaba
a1 kopensko	b1 podzvočno	c1 propelersko	d1 kratkega doleta	e1 potniško
a2 vodno	b2 nadzvočno	c2 turbopropelersko	d2 srednjega doleta	e2 tovorno
a3 kopensko in vodno		c3 reaktivno	d3 dolgega doleta	e3 potniško in tovorno
				e4 večnamensko

Eden večjih projektov, katerega cilj je urejanje besedišča (splošnega) jezika v hierarhične strukture, je WordNet.⁴ Ta semantični leksikon je bil zgrajen na Princetoni univerzi za angleščino, kasneje pa je bila njegova zasnova uporabljena še za izgradnjo sorodnih baz za številne druge evropske jezike, v zadnjem času tudi slovenščino (Fišer 2007). Angleški WordNet je prosto dostopen na spletu za brskanje, celotno bazo s pregledovalnikom vred pa si lahko tudi prenesemo na lastni računalnik.

⁴ <http://wordnet.princeton.edu>

V besedilih se hierarhična razmerja izražajo z določenimi ustaljenimi besednimi zvezami, ki bralcu pomagajo miselno razvrščati informacije in si ustvarjati celotno sliko o predstavljenem predmetnostnem področju. Z vidika metabesedilne organizacije tako lahko prepoznamo tiste stalne dele besedil, s pomočjo katerih besedilo organiziramo, klasificiramo, interpretiramo oz. ocenjujemo propozicijsko vsebino besedila ali se nanjo odzivamo, vendar k njej ne dodajamo kaj novega (Pisanski 2002: 184).

Malaisé in dr. (2007: 20) namesto o besedilnih označevalcih govorijo o opredeljujočih kontekstih (*defining contexts*), ki določeni termin umeščajo v specializirano terminologijo na podlagi razlikovalnih načel. Tako lahko opredeljujoči kontekst poja-
snjuje:

- podobnost s staršem: semantične lastnosti, ki so skupne terminu in njegovemu nadrejenemu pojmu,
- razliko od starša: semantične lastnosti, ki termin ločijo od nadrejenega pojma,
- podobnost s sorojenci: semantične lastnosti, ki so skupne terminu in istorednim pojmom,
- razliko od sorojencev: semantične lastnosti, ki termin ločijo od istorednih pojmov.

Na podlagi podkorpora naravoslovno-tehničnih besedil v korpusu FIDApplus in zgledov v tuji literaturi (Meyer idr.; 1999; Pearson 1998: 174–175) so bili za slovenščino ugotovljeni nekateri tovrstni besedilni elementi (Gorjanc in Vintar 2007):

- povezovanje alternativnih poimenovanj za isti pojem: *ali, ali tudi, imenujemo (tudi), imenovan tudi, sinonim, je sinonim za, znan tudi kot, znan tudi pod imenom, je poimenovan, nosi ime...*
- hierarhično povezovanje pojmov: *je, kot je (na primer), kot je npr., je vrsta, prištevamo med, sodi med, med * sodi, spada med, spada v družino, uvrščamo med, med * uvrščamo, uvrščamo v skupino...*
- povezovanje pojmov v razmerju del — celota: *ima, ima * dele, je iz, je sestavljen iz, vsebuje...*

Spodnji primer je konkordanca iskalnega pogoja *je_vrsta* z omejitvijo na naravoslovnotehnična besedila korpusa Fidapplus (Slika 5). Iz nje lahko razberemo več parov hierarhično povezanih pojmov z relacijo nad-/podpomenskost: *pljuvajoča naočarka* – *pljuvajoča ovratničarka*, *rod Agave* – *sisal*, *skakač* – *Sminthurus viridis* itd.

Slika 5: Konkordanca *je_vrsta* iz naravoslovnotehničnih besedil korpusa Fida-plus

Evropi pojavlja kot plevel. Ameriški dvojnik navadne marjetice je vrsta <i>B. integrifolia</i> .
Iajpogostejši vrsta ostvarjev ob severozahodnih obalah Atlantika je vrsta <i>Limulus polyphemus</i> , ki sega vse do obal Mehike
ki lovi sesalce, ptiče in krastače. Pljuvajoča ovratničarka je vrsta pljuvajoče "naočarke, ki lahko v obrambi razprši
je doma v Evropi, njena ameriška sorodnica pa je vrsta <i>Sorex palustris</i> . Obe vrsti imata dlakave noge in rep
uporabna tekstilna rastlina, doma v Srednji Ameriki. Sisal je vrsta rodu <i>Agave</i> , ki je predstavnik enokaličniške družine agavovk
živali v zgornjem sloju tal. Nekaj skakačev, kot je vrsta <i>Sminthurus viridis</i> , živi na listih nekaterih rastlin, nekatere
, kar je bilo verjetno posledica neke bolezni. Danes je vrsta verjetno že izumrla, čeprav v nepristopni divjini zahodne Tasmanije
Delavci nekaterih termitov, kot je vrsta <i>Odontotermes transvalensis</i> , gojijo glivice na svojih iztrebkih, ki
. Večinoma so manjši kot domači vrabec, čeprav je vrsta <i>Passerella iliaca</i> skoraj tako velika kot drozg. Ti ptiči
Freeware je vrsta programov, ki jih lahko uporabljamo in presnemavamo zastonj.
kar komunikatorju dejansko da uporabnost za poslovno uporabo, je vrsta dodatne programske opreme za varno uporabo in povezavo s poslovnimi
tanko in upogljivo gumijasto opno. Vodni drsalci (to je vrsta žuželk) lahko hodijo po vodi zato, ker je
je predpisati enoten program za vse rejce konj, saj je vrsta programa odvisna od števila živali. Pri rejcu, ki
BELI PARADOXS (Paradoxe Blanc) je vrsta belega vina sorte šardone (chardonnay), ki so
uporabil na svoji sliki. Tudi na levi strani slike je vrsta malih labirintov med hišami.
in bleščečo plast. Žgani lošč (glazura) pa je vrsta steka. Če domneva drži, je skromno lončarstvo izvor

2.2.2 Strokovno specifična razmerja

Na vseh predmetnih področjih pa se med pojmi vzpostavljajo tudi razmerja, ki niso hierarhična. Takšna tematska razmerja opredeljujejo odnose med pojmi, specifične za določeno strokovno področje. Na področju ontoloških aplikacij se takšna razmerja imenujejo *propozicije*, na področju računalniškega jezikoslovja pa *semantične relacije*.

Strokovno specifična razmerja so ključna za urejanje znanja na določenem področju in se jih v zadnjem času skuša tudi intenzivno formalizirati ter zapisovati v baze znanja oziroma ontologije. Če za primer znova vzamemo področje medicine, je v semantičnem omrežju UMLS opredeljenih 54 pomenskih razmerij, specifičnih za medicino, na primer *causes*, *affects*, *complicates* [*povzroča*, *vpliva na*, *povzroča komplikacijo*] itd. Ta razmerja so nato uporabljena za povezovanje 134 semantičnih razredov, ki jih prav tako opredeljuje UMLS, na primer:

Antibiotic causes Disease or Syndrome
Neoplastic Process associated with Mental or Behavioral Dysfunction
Hormone disrupts Gene or Genome

Na ta način so vsi medicinski pojmi, vsebovani v večjezičnem metatezavru UMLS, del semantičnega omrežja. S pomočjo računalniškega označevanja strokovnih besedil

s koncepti in semantičnimi razredi UMLS je bilo zgrajenih že precej naprednih aplikacij, sprva le za poizvedovanje in iskanje ustreznih dokumentov v zbirki Medline, nato pa tudi za odkrivanje novega znanja.

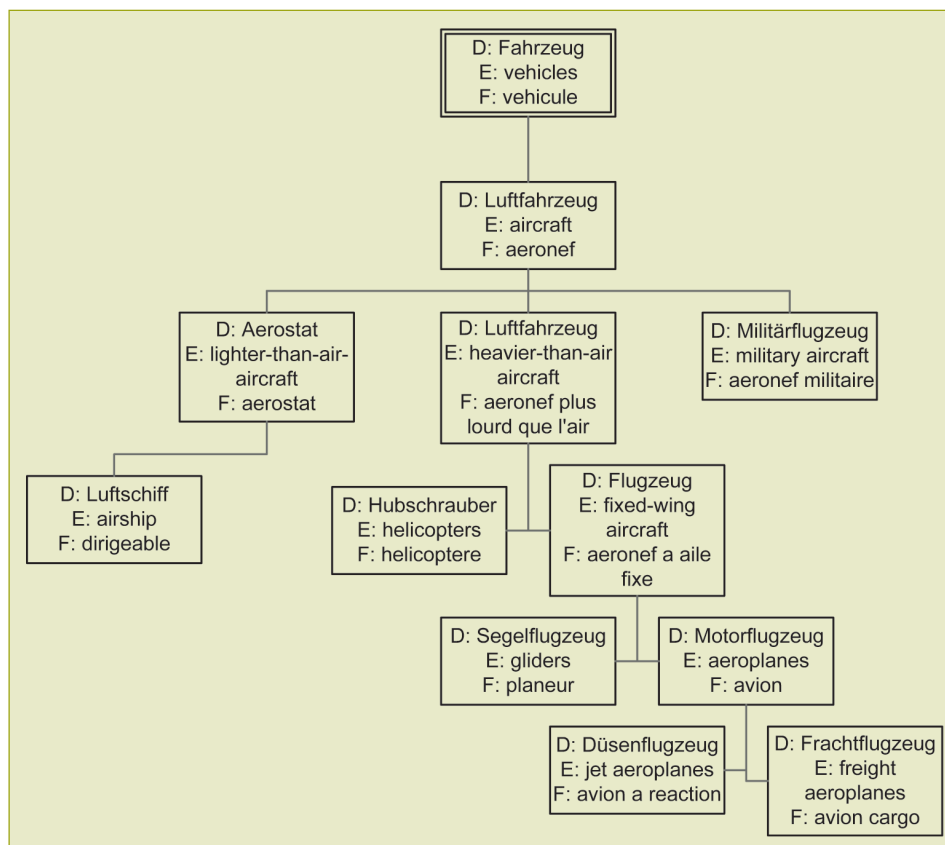
Strokovno specifično razmerje je asimetrični in nehierarhični odnos med najmanj dvema paroma pojmov na določenem strokovnem področju. Prvi korak za opredeljevanje strokovno specifičnih razmerij je zato navadno opredelitev semantičnih razredov, nato pa povezovanje teh razredov med seboj s pomočjo razmerij.

2.3 POJMI V MEDJEZIKOVNIH STIKIH

Univerzalnih miselnih enot, čeprav bi si jih morda želeli, ni; tako so tudi pojmi nedvomno kulturno, verjetno pa tudi jezikovno pogojeni. Z vprašanjem, kako se naše zaznavanje okolja odraža v jeziku in kako – obratno – naš jezikovni sistem vnaprej oblikuje naše miselne strukture in dojemanje sveta, se je ukvarjala množica jezikoslovcev, filozofov, psihologov in biologov skozi stoletja. Če se na tem mestu izognemo razvpitemu mitu o eskimskih izrazih za sneg, pa je bila v prejšnjem stoletju nedvomno precej vplivna Sapir-Whorfova hipoteza (Whorf 1956), ki poudarja vpliv jezikovnega sistema na zaznavanje in strukturiranje resničnosti. Njuna teorija se opira na študije številnih evropskih jezikov v primerjavi z nekaterimi indijanskimi jeziki, ki kažejo na predmetno (oz. samostalniško, statično) naravnost evropskih jezikov ter povsem drugačno, postopkovno usmerjenostjo jezika Indijancev plemena Hopi.

Novi pojmi se prenašajo iz jezika v jezik prek svojih poimenovanj, zato ta sooblikujejo tvorjenje miselne enote. Priporočilo klasične terminološke vede, da večjezikovno terminološko delo vključuje najprej izdelavo pojmovnih sistemov v posameznih jezikih, nato pa njihovo primerjavo in iskanje prevodnih ustreznic, je uporabno le toliko časa, dokler se pojmi kot miselne enote v različnih jezikih pokrivajo. Kadar je pomensko pokrivanje le delno oziroma gre za večje razlike v pojmovnih strukturah, je sicer še vedno naloga terminologije, da te razlike primerno opiše in poišče ravni ustrezanja, vendar pa se nam v takšnih situacijah pogosto zazdeva, da je razmerje med pojmom in poimenovanjem precej bolj intimno, kot ga razlaga pojmovno načelo.

Za primer vzemimo Felberjevo shemo delitve zračnih plovil, iz katere je razvidno, da v angleškem pojmovnem sistemu *aeroplane* ni nadpomenka pojma *glider*, medtem ko v slovenščini *jadravno letalo* vsekakor uvrščamo med *letala*. Primerjava pojmovnih sistemov razkrije, da sl. *letalo* približno ustreza angl. *heavier-than-air aircraft*, in obratno angl. *aeroplane* približno ustreza sl. *motorno letalo*.

Slika 6: Večjezikovno pojmovno polje z izhodiščem angl. *aircraft*

2.4 TERMINOLOGIJA IN SEMANTIKA: OD SLOVARJA DO ONTOLOGIJE

Semantika oziroma pomenoslovje se ukvarja s pomeni leksemov ter razmerji med njimi. Če torej želimo opisati vse možne pomene določene besede, neizogibno sežemo v abstraktni svet pojmov, saj se pomeni besed, stavkov in besedil ustvarjajo prek kognitivnih procesov razumevanja jezika. Očitno se torej tako semantika kot terminologija v prvi vrsti ukvarjata z miselnimi abstrakcijami, torej pomeni in pojmi, ter razmerji med njimi.

Pri tem je razlika med semantiko in terminologijo predvsem na dveh ravneh, in sicer prvič v splošnosti oziroma specializiranosti, drugič pa v sami smeri preučevanja.

Semantika pri svojem ukvarjanju s pomeni ne dela razlik med splošnimi in specializiranimi leksemi, ker pa so slednji pogosto enopomenski, z vidika semantike niso najbolj zanimivi za raziskovanje. Nasprotno pa so za semantiko izredno zanimive vse enote, ki izkazujejo bodisi večpomenskost, se pravi da en jezikovni znak označuje več pomenskih enot, ali nekompozicionalnost, se pravi da iz kombinacije več pomenskih enot ne nastane vsota vsebovanih pomenov, ampak nov pomen.

Razlika med semantiko in terminologijo je tudi v smeri pristopanja k urejanju pomenov. Prva pretežno izhaja iz leksemov in definira pomenska razmerja na podlagi leksikaliziranih enot, se pravi obstoječih jezikovnih prvin. Druga pa, vsaj teoretično, izhaja iz pojmov določene stroke in se v prvi fazi ukvarja z razmerji med njimi, nato pa na tako pridobljeni model področja pripenja jezikovna poimenovanja (in druge reprezentacije).

Za terminologijo, še posebej v večjezikovnem kontekstu, je ukvarjanje s pojmi nujna predstopnja ukvarjanju s termini, saj dobra strukturiranost znanja na nekem področju pripomore k poimenovalni učinkovitosti. Ko pa govorimo o strukturiranju znanja, ne moremo mimo ontologij.

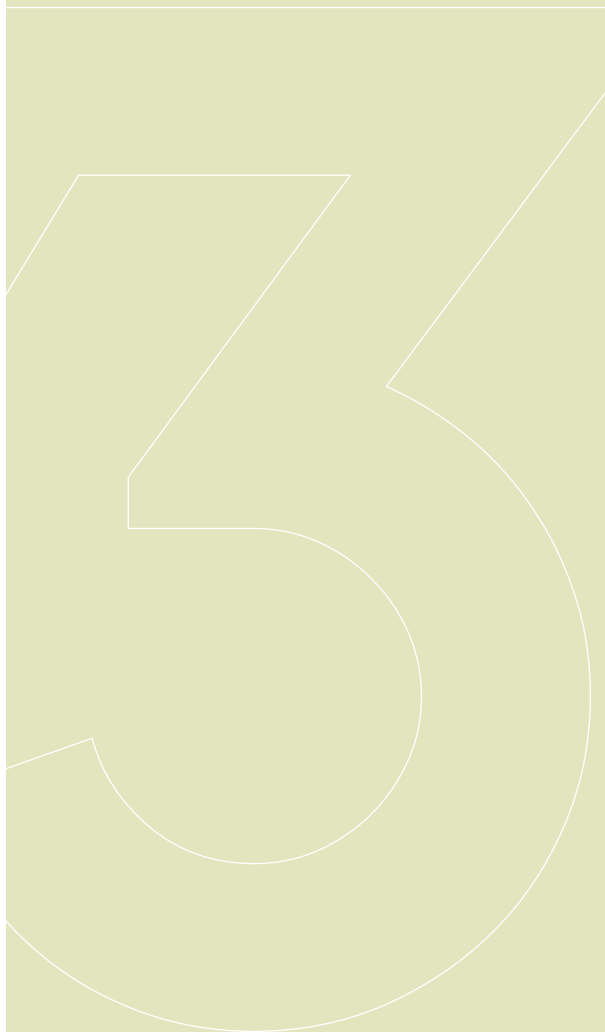
Ontologija je izraz, sposoben iz filozofije, ki na področju računalništva in informatike označuje formalno urejeno strukturo pojmov določenega področja in razmerij med njimi za namene inteligentnih aplikacij. Slednje se tu v najširšem smislu nanašajo na kakršne koli sisteme, ki so zmožni sklepanja oziroma napredne interpretacije podatkov, denimo sisteme za odgovarjanje na vprašanja, semantično podprto poizvedovanje po podatkovnih zbirkah ali odkrivanje znanja.

Ontologija je teoretično jezikovno neodvisna, čeprav se za reprezentacijo pojmov in razmerij v njej uporablja naravni jezik. Tako v praksi obstajajo predvsem ontologije v angleškem jeziku, nekaj jih je večjezičnih, manjšina pa jih za poimenovanje pojmov uporablja katerega od neangleških jezikov. Kot vidimo, je torej terminologija nujen del ontologije, še posebej ker večina ontoloških aplikacij dostopa do podatkov, izraženih v naravnem jeziku.

Spletno mesto OntoSelect⁵ (Buitelaar et al. 2004) je nekakšno samodejno zbirališče obstoječih ontologij za različna področja, kjer je trenutno zabeleženih 1652 prosto dostopnih formalnih ontologij. Žal niti ena od njih ne vsebuje tudi slovenščine, kar je podatek, ob katerem se velja zamisliti. Če bo namreč semantični splet bodočnosti ponujal celo vrsto inteligentnih novih storitev, temelječih na znanju, formaliziranem v ontologijah, nam grozi, da nas bo globalni razvoj tehnologij znanja na široko zaobšel.

⁵ <http://olp.dfki.de/ontoselect>

3 Prepoznavanje in opis terminološkosti



Osnovni problem vsakega ukvarjanja s terminologijo, pa najsi gre za gradnjo terminološke baze ali slovarja, klasifikacijo strokovnega besedila ali strokovni prevod, je odločanje o tem, kaj je termin in kaj ne. V nadaljevanju se tako posvetimo različnim teoretskim pristopom k definiciji termina, nato spregovorimo o oblikah terminov ter njihovih variacijah, nazadnje pa še o uporabniških vidikih terminološkosti.

3.1 KLASIČNE IN PRAGMATSKE DEFINICIJE TERMINA

Po definiciji ISO 1087, ki je v skladu s Felberjevim in Sagerjevim pojmovanjem (Sager 1990), je termin opredeljen kot jezikovno poimenovanje opredeljenega pojma:

term: Designation of a defined concept in a special language by a linguistic expression.

Ob tem Sager (1998/99) ugotavlja še, da se termini po svoji zunanji podobi ali obliki z ničemer ne ločijo od običajnih besed. To je pomembno opažanje, saj iz njega sledi, da ne moremo oblikovati formalnih kriterijev za razlikovanje terminov od neterminov, obenem pa nam daje vedeti — kot bomo nadrobneje videli v nadaljevanju — da je edini možni kriterij za omenjeno razlikovanje funkcija, ali z drugimi besedami specializirana raba.

Pri tem se v klasični terminologiji predpostavlja, da so pojmi znotraj določenega področja opredeljeni in drug od drugega jasno razmejeni, prav tako pa so opredeljena razmerja med njimi. V idealnem svetu je razmerje med terminom in pojmom ena proti ena, se pravi da en termin ne more poimenovati več pojmov, za en pojem pa se sicer lahko pojavi več poimenovanj, vendar je v tem primeru naloga terminologije, da nastopi predpisovalno in izmed obstoječih variant izbere eno.

Resničnost strokovnih jezikov je vse kaj drugega. Obstoječi pojmi se spreminjajo, nastajajo novi, njihova poimenovanja pa zaostajajo za samo predmetnostjo. Poimenovanja za nove pojme se pogosto delno prekrivajo s starimi in nikakor ne drži, da so razmerja med njimi vselej jasno opredeljena. Ker strokovno izrazje ni omejeno zgolj na strokovni in znanstveni diskurz, ampak se delno prekriva s splošnim jezikom in z drugimi področji, je predpisovalno načelo, ki naj bi zagotavljalo enoznačnost, nemogoče udejanjati. Potrebujemo torej drugačna merila terminološkosti, ki nam bodo pomagala ločevati termine od običajnih besed in besednih zvez.

3.2 BESEDILNI PRISTOPI K TERMINOLOŠKOSTI

Pragmajezikoslovni pristop k terminologiji izhaja iz strokovnega besedila, ki naj bi vsebovalo tri kategorije izrazja glede na kriterij strokovne specifičnosti (Hoffmann 1985):

1. strokovno specifično izrazje,
2. splošno strokovno izrazje in
3. splošno izrazje.

V prvo skupino bi torej uvrstili termine, ki sodijo v stroko, ki jo besedilo obravnava, v drugo tiste izraze, ki jih sicer zaznavamo kot strokovne, a si jih delijo številne stroke, in v zadnjo besede splošnega, nestrokovnega jezika.

Kriterij za razvrščanje po teh kategorijah je t.i. *strokovno-specifična referenca*, se pravi za stroko specifični in opredeljeni pomen, ki rabo izraza v stroki ločuje od splošne rabe. Če na primer srečamo izraz *reševanje problema* ali *sklepanje* v besedilu, ki govori o gradnji ekspertnih sistemov z umetno inteligenco, ju lahko na podlagi njune specifične reference uvrstimo med strokovno specifične izraze. Splošnejše strokovne izraze, kot so *sistem*, *kontrola*, *program* pa bi po tej shemi uvrstili v drugo skupino.

Kot pokaže Pearsonova (Pearson 1998), je ta pristop intuitivno sicer naraven, saj izhaja iz besedil, vendar je za potrebe utemeljenega razlikovanja med termini in netermini neuporaben, saj lahko v vmesno skupino splošnega strokovnega izrazja pospravimo vse mejne primere, ne da bi dejansko ugotovili stopnjo specifične reference oziroma terminološkosti. Kriterij razvrščanja je nujno subjektiven, saj temelji na (ne)poznavanju določenih izrazov — nemogoče je namreč jasno ugotoviti, kaj je splošno in kaj ne.

Pearson na podlagi pomanjkljivosti tradicionalnih in pragmatiskih pristopov ugotavlja, da je za ustrezno razlikovanje med termini in netermini potrebno upoštevati vsaj naslednje:

- Ljudje v različnih situacijah različno komuniciramo. Terminološkost je značilnost določenih komunikacijskih situacij, zato je analiza te situacije ključna postavka pri ugotavljanju, ali je jezik določenega diskurza terminološki ali ne.
- O terminološkosti določenega jezikovnega sredstva nam največ pove njegovo sobesedilo, zato so pristopi, ki skušajo značilnosti terminov zaobjeti brez upoštevanja sobesedila, nerelevantni.

V nadaljevanju Pearson terminološkost opredeli kot značilnost specializiranega ali strokovnega jezika (angl. *sublanguage*), se pravi jezika, ki je v rabi v določeni komunikacijski situaciji oziroma diskurzu. Najbolj tipične situacije, kjer se pojavljajo termini, so komunikacija strokovnjak – strokovnjak, strokovnjak – strokovno poučeni, strokovnjak – strokovno nepoučeni, učenec – učitelj, pri čemer ima vsak diskurz svoje značilnosti glede rabe terminologije. V izobraževalnem okolju se denimo poleg terminov uporablja

tudi širok nabor razlagalnih in opredeljujočih sredstev, pogoste so definicije in razlage, ki uporabljajo splošni jezik, visoko specializirani termini pa so razmeroma redki.

Nenazadnje na našo presojo, ali je neko jezikovno sredstvo termin ali ne, precej vpliva tudi poznanost oziroma razširjenost izraza, ali z drugimi besedami, delež oseb v celotni populaciji nekega jezika, ki jim je izraz poznan. Ta kriterij je sicer izrazito subjektiven, po drugi strani pa izrazito intuitiven, kar pomeni, da ga v okviru svojega jezikovnega čuta lahko uporablja prav vsakdo, brez kakršnih koli znanj o načelih terminologije. Na podobno subjektiven način govorci jezika zaznavamo tudi tujost, se pravi ali je določeni izraz prevzet ali ne.

Če nazadnje na terminološkost pogledamo še z vidika prevajalca, ki ga glede na strokovno poznavanje področja, ki ga prevaja, lahko uvrstimo med pollaike, se nam razkrije še dodaten spekter meril terminološkosti. Poleg strokovno specifične reference — ki je, v najslabšem primeru, neuki prevajalec niti ni sposoben prepoznati — se pri prevajanju nedvomno uporablja tudi merilo (ne)poznanosti, posebno vlogo pa igra vidik ustaljenosti v frazeološkem smislu. Ustajene besedne zveze ali *terminološke kolokacije* so za prevajalca pri tvorjenju funkcionalnega strokovnega besedila tako rekoč enako pomembne kot pravi termini, z edino razliko in pomanjkljivostjo, da jih redkeje srečamo v terminografskih priročnikih.

3.3 OBLIKA TERMINOV

Glede na obliko lahko termine razdelimo na več skupin:

- **enobesedni**
 - enostavni: *celica*, *klon*
 - sestavljani: podstava + obrazila: *acikličen*, *polisemija*, *nadzvočni*
podstava + podstava (+ obrazila): *avtocesta*, *ogljikovodik*, *fotosinteza*
- **večbesedni**
 - tvorjeni po različnih skladenjskih vzorcih (glej Tabelo 3)
- **krajšave**
 - kratice
 - začetnice ali akronimi: *SAZU*, *MZT*, *ZDA*, *UNESCO*, *HTML*
 - krnjene besede: *meter* > *m*, *liter* > *l*
 - krnjene zloženke: *bit* (*binary digit*), *Tosama*, *Amex* (*American Stock Exchange*)
 - simboli, formule: *NaCl*, *5a + 3b*
 - okrajšave: *d. d.*, *ekspr.*, *pribl.*

Tipologija krajšav je povzeta po Pravilih Slovenskega pravopisa (Toporišič in dr. 2003), členi 1020-1025. Besedno berljive kratice, ki se pogosto uporabljajo, lahko v pisavi in sklanjanju preidejo v lastna imena (Unesco, Unesca). Okrajšave v slovenščini pišemo s piko, v vezanem besedilu jih, če je le mogoče, razvežemo.

Poseben vidik oblike terminov je njihova besednovrstna oziroma, natančneje, oblikoskladenjska sestava. V nadaljevanju tega razdelka za besedne vrste v slovenščini uporabljamo kratice S za samostalnišnik, G za glagol, P za pridevnik, R za prislov, V za veznik in D za predlog. Večina terminov je samostalniških, se pravi so sestavljeni bodisi iz enega samostalnika bodisi iz besedne zveze, v kateri je jedro samostalnišnik. Močno prevlado samostalniških terminov pokažejo tudi različne statistične analize obstoječih terminoloških zbirk (Logar in Vintar 2008).

Najpogostejši besednovrstni vzorec v Evrotermu je P S, med tipične terminološke vzorce v slovenščini pa prištevamo še S S, P P S, P S S, S D S in druge. Nekaj primerov iz Evrotermu in iSlovarja za vsak besednovrstni vzorec je navedenih v Tabeli 3.

Tabela 3: Besednovrstna sestava slovenskih terminov

Vzorec	Primeri
P S	aromatična snov, fiskalni agent, človekove pravice; brezžična aplikacija, protivlomna naprava
S S	emisija snovi, agent ladje, stalež goved; vir podatkov, zbirka podatkov
S S S	odbor načelnikov štabov; način stanje pripravljenosti
P S S	človekove pravice žensk; dosežena raven storitve
S P S	stalež lovljenih rib; obdelava naravnega jezika
S D S	odpravnina za vdovo(ca), prošnja za azil; izobraževanje na daljavo
P P S	govorno podprta aplikacija, izvajalna podatkovna shramba, predmetno usmerjen jezik
S D S S	Odbor o gibanju blaga
S P S S	določitev kemične sestave snovi

Načeloma lahko termin vsebuje vse besedne vrste, pa tudi dolžina termina ni z ničemer omejena. Še posebej v pravni in politični terminologiji so lahko termini dolgi tudi po deset in več besed, denimo kadar gre za poimenovanja različnih teles (npr. *Odbor Evropskih skupnosti za bombažno in sorodno industrijo (Eurocoton)*, *Odbor pristojnih organov za Direktivo o uporabi genetsko spremenjenih organizmov v zaprtih sistemih*, *Agencija ZN za pomoč in zaposlovanje palestinskih beguncev na Bližnjem*

vzhodu). Ob tem je treba omeniti, da Evroterm kot terminološka baza, ki dejansko nastaja ob prevajanju, vsebuje precej besednih zvez, za katere bi težko trdili, da so poimenovanja določenih jasno opredeljenih pojmov, so pa nedvomno terminološke kolokacije, ki so za prevajalce enako pomembne kot sami termini.

Pričakovana samostalniškost terminov pa je pogosto razlog, da neupravičeno spregledamo termine drugih besednih vrst, še posebej glagole. Ker si tudi pojme zamišljamo kot statične predmetnosti, nam glagoli kot poimenovanja za pojme ne pridejo ravno prvi na misel. V eni novejših študij o (de)terminologizaciji glagolov Žele (2004:78) ugotavlja:

Glagoli so zaradi svoje organizacijske vloge v stavčnih povedih povsem netipična besedna vrsta za termine — večina potencialnih glagololskih terminov je že tako in tako izimenskih.

V nadaljnji obravnavi tudi Žele, podobno kot Pearson, kot merilo za ugotavljanje terminološkosti glagola postavi samo strokovno besedilo, kjer razlikuje med primarno in sekundarno terminološkostjo. Prva je lastnost glagolov z ozko specializiranim pomenjem (*uzimovati [čebele]* čeb., *peskati [ulitke]* teh., *koksati [črni premog]* metal.), druga pa lastnost temeljnih ali pomensko primitivnih glagolov, ki se navadno pojavljajo kot del terminološke besedne zveze (*odpraviti [neznanko]* mat., *odpraviti [pošiljko]* ptt, *odpraviti [vlak]* žel.).

Če se ozremo naokoli, glagole vsebuje precej terminografskih del novejšega časa, tako denimo Vojaški slovar (Korošec 1998), iSlovar (Puc in Erjavec 2006), Planinski terminološki slovar (Klinar 2004) in drugi. Slednji ne vsebuje le nedoločniških glagol-skih oblik (*plezati, sestopati*), ampak tudi velelniške in vprašalne oblike, medmete in prislove kot plezalne ukaze (*Dol! Sem! Hej!*). Razlaga pri spodnjem primeru je dokaz, da gre vsekakor za terminološki izraz.

Gréš? vprašanje varujočega spodnjemu ali napredujočemu soplezalcu v navezi, ali bo začel plezati

Terminološke prislove najdemo tudi na področju glasbene terminologije, denimo *allegro, andante con motto, forte, rubato* itd., nekateri od teh pa lahko privzamejo tudi samostalniško rabo, npr.

Želja po interpretativnem osmišljanju prav vsakega tona je v zaključnem allegrettu namesto pričakovanega (in v Beethovnovi partituri zapisanega) ritmičnega ostinata prinesla pohitevanja in umirjanja, povsem odsotno igro in nenadne izbruhe.

Dnevnik, 12.2.2002 (vir: FidaPlus)

3.4 TERMINOLOŠKE VARIACIJE

Čeprav se večbesedna terminološka enota navadno opredeljuje kot ustaljena ali celo fiksna, se to tradicionalno pojmovanje termina kot jezikovno nespremenljive enote izkaže za neustrezno, saj opazovanje terminov v strokovnih besedilih priča o tem, da se strokovno izrazje v rabi približuje značilnostim splošnojezikovnih leksikalnih enot. To pomeni, da se večbesedni strokovni izrazi v besedilih obnašajo podobno kot večbesedne fraze splošnega jezika in tako prehajajo skozi različne postopke oblikoslovnih, skladenjskih in pomenskih transformacij.

Ker so terminološke študije zelo dolgo temeljile na jezikoslovnih intuicijah terminologov in ročnih analizah besedil, je bilo vprašanje terminoloških variacij dolgo obrobnege pomena, saj se jih je večinoma pojmovalo kot »neljube odklone« od ustaljenih poimenovanj. Z besedilno motiviranimi metodami pa se terminološke variacije postavljajo v ospredje, saj jih na eni strani lažje odkrivamo — nenazadnje nam že običajna konkordanca jedra terminološke zveze razkrije vse njene variacije —, na drugi strani pa nam variabilnost terminologije otežuje samodejno prepoznavanje v programskih aplikacijah za luščenje izrazja.

S terminološkimi variacijami se je s stališča računalniške obdelave prvi sistematično ukvarjal Jacquemin (2001), ki poda naslednjo definicijo terminološke variacije:

Oblikoslovna, skladenjska ali pomenska variacija je transformacija kanonične večbesedne oblike termina, ki zadostuje naslednjim kriterijem:

- *Vsebinske sestavine termina se ob transformaciji ohranijo oziroma se preobrazijo v oblikoslovno ali pomensko sorodne besede. Tako je na primer recognized neural cells prava variacija termina Cell recognition, kajti ohrani se sestavina cell, sestavina recognition pa se preobrazi v oblikoslovno sorodno besedo recognize. V postopku transformacije lahko izginejo oziroma se nadomestijo zgolj »prazne« besede, kot so predlogi, vezniki ali členi.*
- *Med vsebinskimi sestavinami je možna oblikoslovna ali pomenska sorodnost.*
- *Variacije, ki ne vključujejo razmerja besedotvorne preobrazbe niti razmerja pomenske preobrazbe, se imenujejo skladenjske variacije. Tako je comprehension of language skladenjska variacija Language comprehension.*
- *Variacije, ki nastanejo s postopkom besedotvorne preobrazbe, so oblikoslovne variacije. Tako je determine the structures primer oblikoslovne variacije Structure determination, saj sta determine in determination oblikoslovno sorodni besedi.*
- *Variacije, ki vključujejo pomensko razmerje, se imenujejo pomenske variacije. Tako je speech comprehension pomenska variacija language comprehension, kajti speech in language sta pomensko sorodna leksema.*

Transformacija lahko spremeni besedni red večbesednega termina, prav tako se lahko skozi njeno ustavi nove besede, vendar morajo odvisnostna razmerja med sestavinami ostati nespremenjena.

Variacija ne sme vsebovati izvirnega termina ali katere od njegovih pregibnih oblik. Tako mean arterial pressure ni variacija arterial pressure, ker je drugi termin gnezden v prvem.

Jacquemin 2001: 4-5.

Variabilnost izrazja se razlikuje od področja do področja, odvisna pa je tudi od besedilne vrste in registra. Uveljavljena strokovna področja, kjer je izrazje v veliki meri standardizirano in vsebovano v normativnih terminoloških priročnikih, naj bi bila manj nagnjena k variacijam, čeprav konkretni primeri pogosto tudi tu kažejo drugače. Izsek iz enega največjih terminoloških tezavrov na svetu za področje medicine UMLS (Unified Medical Language System) kaže nekaj uradnih variacij, ki jih najdemo za izraz *legal abortion*. Po pojmovnem identifikatorju na začetku vrstice (C0000812) vemo, da gre pri vseh za isti pojem.

```
C0000812|ENG|P|L0000812|VW|S0494795|Legal abortion|3|
C0000812|ENG|P|L0000812|VW|S1424857|legal abortion|0|
C0000812|ENG|S|L0183339|PF|S0247930|Legally induced abortion|0|
C0000812|ENG|S|L0183339|VO|S0675645|Legally induced abortion NOS|3|
C0000812|ENG|S|L0487178|PF|S0686453|Medical abortion|3|
C0000812|ENG|S|L0662916|PF|S0929847|Legal abortion unspecified|3|
C0000812|ENG|S|L0662916|VO|S1054769|Unspecified legal abortion NOS|3|
C0000812|ENG|S|L1348135|PF|S1594013|Legally induced abortion - TOP|3|
```

Iz Jacqueminove razprave je razvidno, da je klasično pojmovanje termina kot ustaljene leksikalne enote pri korpusnih terminoloških metodah preživeto, saj se v besedilih termini obnašajo podobno kot leksikalne enote splošnega jezika. Ker se razumevanje strokovnega besedila, tako kot pri vseh vrstah besedil, ustvarja na pojmovni, se pravi miselni ravni, je strokovna poimenovanja mogoče variirati, spreminjati in okrajševati, ne da bi s tem ogrozili koherentnost besedila.

Nekatere variacije je mogoče formalno opisati in jih s tem predvideti, še posebej to velja za skladišne variacije. Jacquemin denimo opisuje sistem za luščenje in razpoznavanje variacij FASTR, ki deluje na podlagi skladišnih pravil za strukturo terminov in transformacijskih pravil za tvorjenje variacij. Tako lahko napišemo pravilo za iskanje atributivnih priredij, ki predstavljajo variacijo skladiškega vzorca pridevnik (P) + samostalni (S), na primer *umbilical and cerebral artery* kot variacija kanoničnih oblik *umbilical artery* in *cerebral artery*. Morda je na prvi pogled videti, da bi takih pravil potrebovali le nekaj, pa bi že zajeli velik del varia-

cij, vendar stvar ni tako enostavna. Če namreč tako pravilo definiramo eksplicitno kot

$$P_1 \text{ and } P_2 + S = P_1 + S \text{ and } P_2 + S,$$

smo s tem opisali le eno od številnih možnih variacij levega prilastka, saj so možne na primer tudi tvorbe *umbilical and middle cerebral artery*; *umbilical, cerebral or carotid artery*; *umbilical or other non-dorsal artery* itd. Transformacijska pravila so zato formalizirana in prirejena slovnična pravila, ki opisujejo teoretično možne strukturne vzorce določenega jezika.

Če imamo na primer za angleščino definirano pravilo, da je termine oblike

$$SZ_1 + \text{of} + SZ_2$$

kjer je SZ samostalniška zveza, možno pretvoriti v kanonično obliko

$$SZ_2 SZ_1$$

nam sistem lahko samodejno predlaga pretvorbo *replacement of the steam generator* v *steam generator replacement*.

V slovenščini najdemo številne primere variacij. Naslednjih nekaj primerov je iz korpusa besedil s področja jedrske tehnike, kjer v terminološkem gnezdu ob besedi *reaktor* zasledimo naslednje variacije:

reaktorski hladilni sistem
sistem reaktorskega hladila
sistem reaktorskega hladiva

cev reaktorskega hladiva
cev sistema reaktorskega hladila
cevovod reaktorskega hladilnega sistema
cevni vod reaktorskega hladilnega sistema

reaktorska stavba
stavba reaktorja
reaktorska zgradba

Ob poskusu, da bi primere slovenskih variacij klasificirali v skladu z zgornjo definicijo, povzeto po Jacqueminu, pa naletimo na nekaj težav. Če vzamemo za primer

obliki *reaktorski hladilni sistem* in *sistem reaktorskega hladila*, se težko odločimo, ali gre za oblikoslovno ali skladenjsko variacijo, saj se v tem in v številnih drugih primerih ta dva tipa med seboj prekrivata. Videti je, da se Jacqueminova definicija bolj prilega angleškemu in francoskemu jeziku, kjer skladenjska pretvorba ne vključuje nujno tudi oblikoslovne. Nasprotno pa se v jezikih, ki skladenjska razmerja izražajo pretežno s sklonskimi obrazili, ta dva tipa preobrazbe nujno zgodita istočasno, zato se postavlja vprašanje smiselnosti takega razlikovanja med njima.

Problematično je tudi uvrščanje pomensko sorodnih terminov med variacije, še posebej glede na primer *language comprehension* in *speech comprehension*, ki ga navaja Jacquemin. Temeljna lastnost variacij je namreč predpostavka, da se različne jezikovne oblike termina nanašajo na isti pojem, pri pomenskih variacijah pa gre pogosto tudi za odklon od pomenskega polja pojma. Definicijo pomenske variacije bi zato morali nujno dopolniti z zahtevo, da je razmerje med transformiranimi sestavinama sopomenskost. Primer pomenske variacije iz našega korpusa bi lahko bil *reaktorska stavba* in *reaktorska zgradba*.

Ob primeru *sistem reaktorskega hladila* in *sistem reaktorskega hladiva* se postavlja vprašanje, ali gre za oblikoslovno, pomensko ali kako drugo variacijo. Tu gre pravzaprav za pojav morfemske variacije, kjer sta morfema *-ilo* in *-ivo* vzporedna in sinonimna. Če je tak morfološki par produktiven, lahko tvori tudi druge vzporedne oblike, na primer *mazilo* in *mazivo*. Terminološka variacija je tako v tem primeru posledica morfemske variacije, ki se prenese na termin.

Ker je podobne pare vzporednih in potencialno sinonimnih morfemov mogoče zaslediti tudi v drugih jezikih, na primer *-ic* in *-ical* v angleščini, se nam zdi smiselno ta tip variacije obravnavati kot posebno kategorijo.

Variacije se pojavljajo tudi kot *elipse* pri anaforičnih pomenskoreferenčnih nitih v besedilu. Tako se večbesedni termin ob prvi omembi pojavi v kanonični obliki, na primer *dokumentacija za pridobitev gradbenega dovoljenja*, ob nadaljnjih omembah pa v skrajšani različici *gradbena dokumentacija*. Ta tip variacije je verjetno tudi najbolj razširjen in zajema tudi pojavljanje akronimov namesto izpisane različice termina.

Očitno je, da se tipi variacij med seboj prekrivajo in da se pogosto zgodi transformacija na številnih ravneh hkrati. Pri avtomatski obdelavi, ki naj bi vključevala razpoznavanje, morda pa tudi predvidevanje terminov in njihovih variacij, potrebujemo natančna transformacijska pravila, ki vključujejo polno skladenjsko analizo dane strukture ter slovnična pravila za njene možne pretvorbe.

Številne pretvorbe v slovenščini pa vključujejo prehod iz ene besedne vrste v drugo, kar je za avtomatsko obdelavo večji problem. Da bi lahko ustrezno razpoznavali in predvidevali pretvorbe tipa *strdek v krvi* → *krvni strdek*, bi potrebovali oblikoskladenjski leksikon, ki bi vseboval tudi besedotvorne izpeljanke iz določenega korena. Večina trenutno dostopnih leksikonov slovenskih besednih oblik navaja zgolj oblike in oznake iste paradigme, se pravi v okviru ene besedne vrste, ne povezuje pa med seboj besedotvorno sorodnih lem. To vrzel bo morda v kratkem zapolnil slovenski semantični leksikon SloWNet (Fišer 2007).

3.5 TERMINOLOŠKOST S STALIŠČA UPORABNIKOV

Podobno kot se jezik ne tiče le jezikoslovcev temveč vseh njegovih uporabnikov, je tudi terminologija področje, s katerim nimajo opraviti le terminologi. Strokovno izrazje se pogosto prepleta s splošnim jezikom in se tako uporablja tudi pri nestrokovnih vrstah komunikacije, vendar takšna raba ni v središču naše pozornosti. Tu se želimo predvsem usmeriti k poklicnim profilom, pri katerih sodi raba strokovnega izrazja med njihove poklicne naloge oziroma funkcije. Opredelimo lahko naslednje skupine uporabnikov terminologije:

Strokovnjaki

Ti potrebujejo strokovno izrazje pri svojem delu za sporočanje strokovnih vsebin. Pogosto sporočajo v več kot enem jeziku.

Dokumentalisti

S tem izrazom označujemo z bibliotekarstvom in arhiviranjem povezano dejavnost označevanja dokumentov s ključnimi besedami ali klasifikatorji, ki se kasneje uporabljajo za iskanje ali priklic iz zbirk dokumentov.

Prevajalci

Ti se s strokovnim izrazjem srečujejo v okviru prevajanja, se pravi ustreznega prenašanja specialističnih leksikalnih enot v druge jezike.

Terminologi

Njihova naloga je predvsem zbiranje in vrednotenje strokovnega izrazja za namene terminografskih del, pa tudi soodločanje pri oblikovanju izrazja.

Že iz tega strnjenega opisa je razvidno, da so načini uporabe terminologije pri teh štirih skupinah zelo različni. Neposredna posledica tega je, da so si pojmovanja vprašanja, kaj je termin, precej nasprotujoča, saj vsaka skupina terminološki inventar obravnava drugače.

Strokovnjakom je najpomembnejša učinkovitost komunikacije, zato se s termini ukvarjajo le skozi njihov pomen. Zanje je gospodarnost izrazja pogosto pomembnejša od jezikovne neoporečnosti, zato se lažje sprijaznijo s prevzetimi izrazi ali akronimi, če so ti bolj funkcionalni. Po drugi strani strokovnjaki številne besede splošnega jezika uporabljajo s specifičnim pomenom, ki ga ostale skupine pri svojem delu zlahka spregledajo. To je tudi pogost vir strokovnih napak pri prevajanju, saj se prevajalec zaradi znane besede ne osredotoči na njen poseben pomen.

Za **dokumentaliste** je strokovno izrazje nabor označevalcev, kjer je posebej pomembno, da so ti med seboj pomensko urejeni v določeno označevalno shemo. Pri odločanju, kateri izrazi najboljše povzamejo besedilo, je glavni kriterij učinkovitost v iskalni situaciji; izbrati je torej treba besede in fraze, ki jih bo iskalec najverjetneje uporabil za dostopanje do v besedilu vsebovanih informacij. Ti izrazi so navadno kratki, le izjemoma več kot dvobesedni, skoraj izključno samostalniške zveze.

Prevajalci zaradi narave svojega dela termine na nek način povezujejo s posebno vrsto »prevajalskih pasti«, se pravi s tistimi leksikalnimi enotami, ki so za funkcionalnost ciljnega besedila še posebej pomembne. Poznavanje prevodnih ustreznici samih terminov zanje ne zadostuje, saj morajo za ustrezen prevod prenesti pomen v celoti. Smiselni prevod je mogoče zagotoviti le s podrobnim poznavanjem stroke in značilnosti njenega žargona. Za prevajalce so tako termini vse strokovno specifične enote, ki vključujejo tudi kolokacije, posebne krajšave in simbole. Po drugi strani prevajalcev ne zanima tisti del izrazja, ki je splošno znan in pri katerem ne pričakujejo prevajalskih težav.

Terminologi so tista skupina, ki se še najbolj poglobljeno ukvarja s celotnim terminološkim inventarjem določene stroke, tako z jezikoslovnega kot s pomenskega vidika. Ker je poglavitni namen terminološkega dela izdelava terminografskih priročnikov, je v interesu terminologov, da med termine uvrstijo čimveč strokovno specifičnih leksikalnih enot. Obenem je za ustrezno terminografsko obdelavo potrebno razjasniti primere večpomenskosti ali terminoloških variacij, zato je terminolog navadno neprestano v stiku s strokovnjaki področja. Za razliko od prevajalcev pa se terminologi navadno manj posvečajo besedilnemu vidiku, se pravi rabi izrazja v kontekstu. To odražajo tudi številni terminološki slovarji in priročniki, ki so zato za prevajalce bistveno manj uporabni.

Razlike med temi štirimi skupinami zelo nazorno pokaže raziskava, ki jo je izvedla Estopà Bagot (1999) v okviru svoje doktorske disertacije. Da bi empirično ugotovila, kako vsaka od omenjenih uporabniških skupin pojmuje termine, je za vsako zbrala

tri strokovno usposobljene predstavnike, ki so dobili nalogo, da v izbranem besedilu s področja medicine označijo strokovne izraze. Glede na naravo njihovega dela so dobili tudi nekoliko različna navodila (Estopà Bagot 1999: 371).

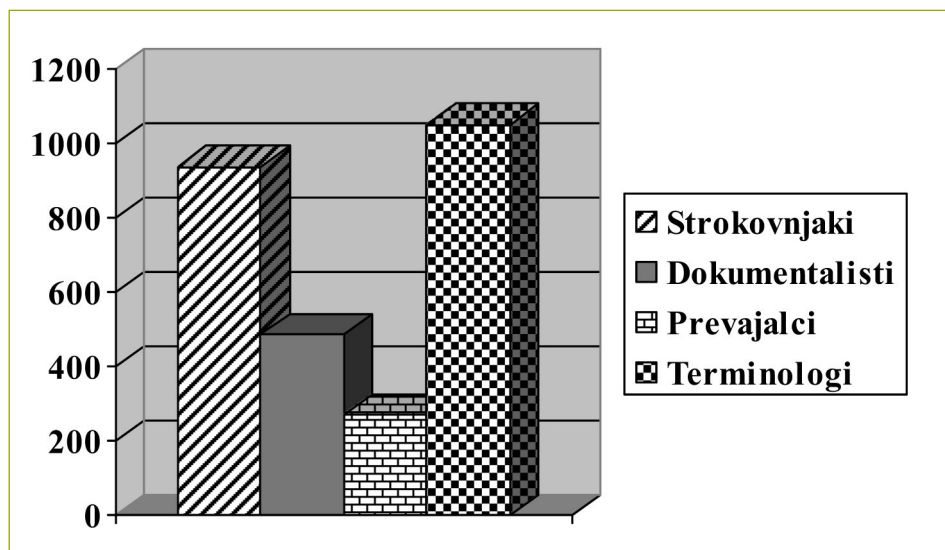
- Strokovnjaki s področja medicine: »Označite vse specializirane leksikalne enote, ki po vašem mnenju označujejo pojme, specifične za medicino.«
- Dokumentalisti: »Označite vse specializirane leksikalne enote, ki bi lahko služile za klasifikacijo oziroma indeksacijo dokumenta.«
- Prevajalci: »Označite vse specializirane leksikalne enote, ki po vašem mnenju otežujejo prevod besedila.«
- Terminografi: »Označite vse specializirane leksikalne enote, ki bi jih vključili v medicinski terminološki slovar.«

Rezultati preskusa so izredno zanimivi, saj se že število izbranih leksikalnih enot po skupinah močno razlikuje (Slika 7). Tako so medicinski strovnjaki označili 938 izrazov, dokumentalisti 486, prevajalci 270 in terminografi 1052. Avtorica študije razlike tolmači s specifikami uporabniških potreb. Tako največ izrazja izberejo terminografi, ki sicer niso področni strokovnjaki, vendar se jim zdi potencialno skoraj vsaka besedna zveza specifična za medicino. Zaradi zavedanja, da imajo lahko tudi običajne besede specifičen pomen, ki bi ga bilo treba v slovarju pojasniti, so pri označevanju najbolj »radodarni«. Tako je morda tudi spodbudno, da se njihov izbor še v največji meri pokriva z izborom strokovnjakov.

Dokumentalisti so označili skoraj pol manj izrazov, med temi tudi skoraj ni bilo drugih besednih vrst kot samostalnikov. Rezultat je pričakovan, saj moramo biti pri označevanju dokumentov »varčnejši« — naloga dokumentalistov je s čim manj izrazi čim bolj podrobno podati vsebino dokumenta. Zanimiv podatek je morda še, da je ta skupina edina označevala tudi lastna imena, ki imajo pri iskanju dokumentov vsekakor precejšnjo vlogo, čeprav jih ne pojmujejo kot strokovne izraze.

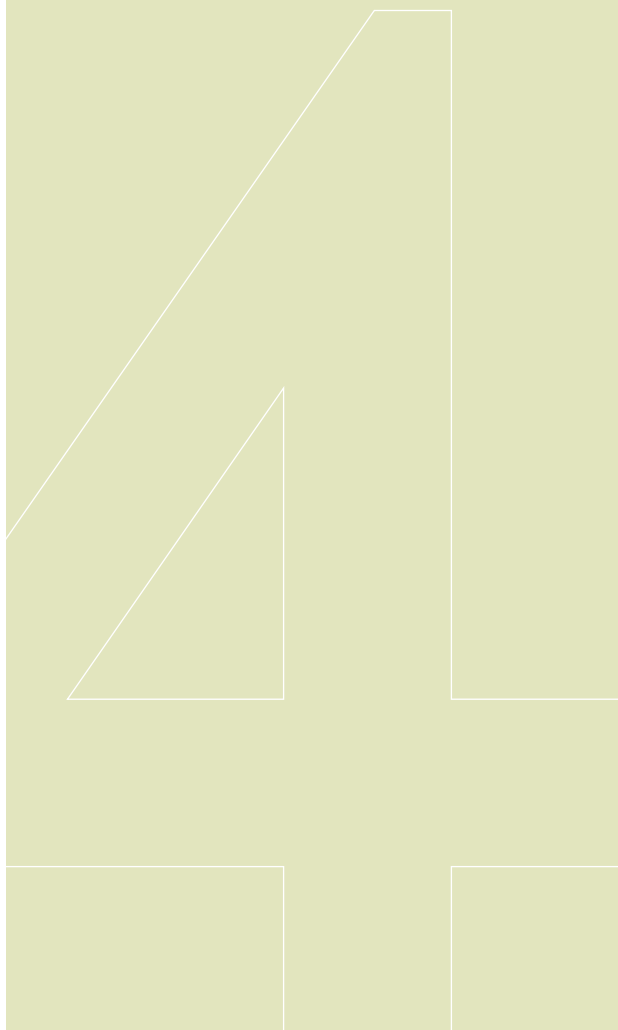
Še najmanj izrazov so označili prevajalci. Razlog za to je verjetno tudi v navodilu, ki je njihovo pozornost načrtno usmerilo v »težavne« enote, ni pa posebej od njih zahtevalo prepoznavanja vseh terminoloških enot. A kljub tej okoliščini rezultat kaže, da so prevajalci verjetno premalo pozorni na strokovno specifičnost splošno znanih besed, ki jih je na področju medicine veliko.

Slika 7: Število terminov, ki jih je označila posamezna skupina



Posebej poučna je ugotovitev, da je presečna množica vseh štirih skupin, tj. izrazi, ki so jih označile vse skupine, obsegala le 119 enot ali približno 10 odstotkov vseh označenih terminov. Tu se že kar šokantno pokaže problematičnost definicije termina in neustreznost posplošujočih pristopov, ki ne upoštevajo uporabniškega vidika. Ob teh rezultatih se umestno postavlja vprašanje, komu naj bi bili terminološki priročniki namenjeni, če so potrebe posameznih uporabniških profilov tako različne. Primerno upoštevanje tega vidika med drugim pomeni, da vsak posamezni glosarski projekt vnaprej definira ciljne skupine uporabnikov ter v skladu s tem zasnuje tudi terminografsko metodologijo.

4 Terminotvorje



Strokovna področja se spreminjajo in razvijajo, z novimi tehnologijami in znanji nastajajo nova, pa tudi na že uveljavljenih področjih se strokovno izrazje neprestano posodablja in dopolnjuje. Nova poimenovanja nastajajo po različnih poteh, pomemben dejavnik pri tem pa je, ali gre za povsem novo predmetnost ali se tvori le poimenovanje v določenem jeziku pod vplivom že obstoječega poimenovanja v tujem jeziku.

4.1 PREVZEMANJE

Prevzemanje oziroma sposojanje iz drugih jezikov navajamo kot prvi terminotvorni proces, saj se dejansko navadno zgodi v prvi terminotvorni fazi, se pravi ko nastopi potreba po komunikaciji v zvezi z novo predmetnostjo, za katero slovensko poimenovanje (še) ne obstaja. Ta faza je lahko tako bežna, da zajema le govorno rabo, v pisno pa sposojenka niti ne pride, lahko pa seveda predstavlja tudi zaključno terminotvorno fazo, ko se sposojeni izraz usidra v jezik, se govorno prilagodi in tvori besedno družino.

V literaturi je opisanih več vrst prevzemanja, denimo znotraj- in zunajjezikovno, kjer je prvo prehajanje terminov iz splošnega jezika v strokovni jezik po načelu metafore in metonimije (*vesoljska ladja*, *deskanje po spletu*, *podatkovna avtocesta*), drugo pa sposojanje morfemov, besed ali besednih zvez iz drugih jezikov. Pri slednjem nadalje razlikujemo med prevzemanjem iz klasičnih jezikov latinščine in grščine ter prevzemanjem iz sodobnih jezikov.

V nadaljnji razpravi se posebej posvečamo slednjemu, saj je prevzemanje iz sodobnih jezikov najbolj produktiven terminotvorni proces, ki je tesno povezan z zgodovinskimi, kulturnimi in ekonomskimi dejavniki. Pri prevzemanju določenega izraza iz enega jezika v drugega gre skoraj vedno za neenakovredno razmerje med jezikom dajalcem in jezikom prejemnikom. Tako je v današnjem času jezik dajalec največkrat angleščina, ker je angleškogovoreče območje sveta ekonomsko in kulturno privilegirano ter iz njega izhaja največ tehnoloških in drugih novosti. Obenem uživa angleščina status *lingue france*, se pravi sporazumevalnega jezika za številna jezikovno nehomogena okolja.

Seveda je na določenih strokovnih področjih lahko tudi kak drug jezik v vlogi dominantnega dajalca. Na področju kulinarike si denimo tudi angleščina ogromno izrazov sposoja iz francoščine in italijanščine, saj v tem primeru temeljno znanje in vrhunski dosežki ne izvirajo iz angleškogovorečega prostora.

Postopek prevzemanja se navadno dogaja v več stopnjah (Bokal 1998):

1. citatni prevzem (*manager*)
2. govorno prilagajanje (*manager [menedžer]*)
3. pisno prilagajanje (*menedžer, favl, dizajn, koktejl,...*)
4. oblikovanje besedne družine s slovenskimi obrazili (*menedžerski, menedžerka,...*)
5. morebitni pojav slovenske sopomenke (*ravnatelj*)
6. opustitev prevzetega izraza

S prvo stopnjo se navadno hkrati zgodi tudi druga, saj Slovenci tuje besede v domačem okolju govorno prilagajamo in izgovarjamo čim bližje slovenskim fonemom. Pri zgodnjih pojavitvah je citatni prevzem pogosto oblikovno ali skladenjsko zaznamovan s poševnim tiskom ali navednicami, takoj ob njem pa se pojavi razlaga, denimo:

Med urjenjem usmerjajo podjetja svojo pozornost na tako imenovani »team building«, torej na graditev in krepitev timskega duha ali na izboljšanje motivacije.

Delo, 26. julija 1999, vir: Nova beseda

Tudi pisno prilagajanje lahko poteka po korakih. Spodnja tabela kaže število zadetkov za besedo *wellness* in njene (delno) podomačene izpeljanke v korpusu FIDAplus. Zagotovo je vsaj del zabeleženih primerov prilagojene pisave pripisati nepoučenosti piscev.

wellness	1303
wellnes	117
welness	64
velnes	33
welnes	2

Če je motivacija za prevzem skoraj vedno leksikalna praznina, ki nastopi ob pojavu nove predmetnosti, pa so razlogi za ohranitev prevzete besede zelo raznoliki. Načeloma namreč za večino strok lahko trdimo, da se trudijo uporabljati domače izrazje povsod, kjer je to mogoče, se pravi kadar obstaja primerna ustreznica. Poseben status v tem okviru ima medicina, kjer raba terminov latinskega izvora pripomore k enoznačnosti in hkrati omogoča lažjo mednarodno izmenjavo.⁶ Prav tako se načeloma

⁶ Z neslovenskostjo medicinskega izraza se še posebej kritično ukvarja Toporišič (1968/2006: 449–461), ki "nepotrebno tujčevanje" pripisuje vsesplošni jezikovni lagodnosti zdravnikov, pa tudi potrebi po dajanju vtisa učenosti oziroma "poskrivnostanju preprostih resnic našega telesa in njegovih stanj". Kot je razbrati iz spremne polemike, medicinska stroka na te očitke odgovarja predvsem z argumentom, da je z uporabo izrazov latinskega izvora medicinski jezik eksaktnejši, po drugi strani pa je takšna raba omejena na znanstvena medicinska besedila, medtem ko se v strokovnih in poljudnih besedilih izrazje prilagaja laiku v smislu rabe pretežno slovenskih izrazov, kjer ti obstajajo.

ne prevajajo kratična poimenovanja različnih strokovnih in upravnih teles EU, razen kjer kratica preide v splošno nestrokovno rabo. Tudi za splošni jezik velja, da izrazi, ki so močno zaznamovani z izhodiščno kulturo, navadno ne dobijo slovenske ustreznice (*pizza, tofu, jazz, citroën,...*).

Tudi na področjih, kjer je opazna težnja po ustvarjanju domačega izrazja, se za določene pojme ohrani prevzeti izraz. Včasih bi bila morebitna slovenska ustreznica bistveno daljša, zato tako zelo ogroža gospodarnost jezika, da se strokovna raba odloči za prevzeti izraz. V drugih primerih se zgodi, da bi bil podomačeni (kalkirani ali novotvorjeni) izraz dvoumen, zato stroka raje ohrani prevzeti izraz (bot. *blanket barje*). Pri številnih izrazih je motivacija za prevzem trženjske narave, sploh kadar gre za izdelek ali storitev, ki je naprodaj (*wellness, team building*).

Da se prevzeta beseda uspešno včleni v jezikovni sistem, naj bi izpolnjevala čim več od spodnjih kriterijev (prim. Bokal 1998):

- prevzeta beseda je potrebna, zapolnjuje leksikalno praznino,
- se uporablja,
- vzpostavlja odnos do drugih enot v pojmovnem sistemu,
- je besedotvorno razvejana in daje podlago za različne pomenske in skladenske izpeljanke,
- je primerno dolga,
- je lahko izgovorljiva.

Številne prevzete besede iz angleščine, še posebej tiste s končnicami *-ness, -ing* ipd., so nerodne že pri sklanjanju, še bolj pa pri tvorbi pridevniških izpeljank (*bodybuilding – bodybuildinški?, wellness – wellneški?*). V pogovornem jeziku zato končnico *-ing* pogosto nadomeščamo s slovensko glagolniško pripono *-anje* (*outsourcing – autsorsanje, downloading – daunloudanje*). Kaže torej, da od naštetih kriterijev nekateri ne vplivajo bistveno na ohranitev prevzete besede v jeziku, drugi — denimo dolžina — pa odločilno.

4.1.1 Izlastnoimensko izrazje

Poseben primer prevzemanja terminov so izlastnoimenski termini; izrazi, ki jih tvorimo iz lastnih imen. Takšnim izrazom pravimo tudi *eponimi*. Posebnost lastnih imen v primerjavi z občnimi je predvsem, da je njihova vloga v jeziku predvsem razločevalna in označevalna, pomena pa bodisi nimajo bodisi je ta potisnjen v ozadje. Košmrlj-Levačič (1998) eponime deli na dve skupini:

1. termini, tvorjeni iz osebnih imen
 - iz rojstnih osebnih imen: *martinček* 'majhen kuščar', *aleksandrit* 'različek hrizoberila' (*geologija*)
 - iz priimkov, navadno raziskovalcev in znanstvenikov, zaslužnih za nova spoznanja: *tesla* 'enota za gostoto magnetnega polja', *Newtonovi zakoni*, *Dopplerjev pojav*, *Geiger-Müllerjev števec*, *rentgenska cev*, *hladnikov grintavec*
2. termini, tvorjeni iz zemljepisnih imen
 - iz krajevnih imen: *petovio* 'sorta orehov' (Ptuj), *londonski peping* 'sorta jabolka'
 - iz imen držav: *grška jelka*, *kanadska gos*, *kalifornij* 'kemični element'
 - iz imen celin: *američij* in *evropij* 'kemična elementa', *avstralski lovor*, *afriška kotorna* 'ptica iz poddružine fazanov'

Če se eponim povsem terminologizira, lahko preide v občno ime. Takšni izrazi se pišejo z malo začetnico. Sem sodijo tudi poimenovanja za merske enote: *hertz*, *joule*, *newton*, *pascal*, *kelvin* itd. Po Odredbi o merskih enotah naj bi imele prednost v celoti poslovenjene oblike: *herc*, *džul*, *njuten*, *paskal*, *vat* itd., vendar se raba v naravoslovnih znanostih pretežno drži citatnega zapisa.

4.1.2 Kalkiranje

Kalkiranje je terminotvorni postopek, kjer slovensko ustreznico tvorimo neposredno po tujejezični predlogi, ali drugače rečeno, kadar prevzeti izraz dobesedno, včasih celo po posameznih morfemih, prevedemo (npr. *internet* v *med-mrežje*, *viktimologija* v *žrtv-o-slovje*, *escape character* v *ubežni znak*). Zaradi svoje povezanosti s procesom prevzemanja ga obravnavamo v tem razdelku.

Ker naj bi se s kalkiranjem tudi v lastni jezik vnašalo tuje videnje sveta, čeprav je izraz dejansko domač, se pri slovenskih jezikoslovcih pogosto zasledi odklonilen odnos do te oblike prevzemanja. Po drugi strani pa kalka sploh ne bomo opazili, če tujega izraza ne poznamo ali preprosto nismo pozorni (več o tem Vidovič Muha 2000 in Logar 2005).

Kadar je prevedeno poimenovanje tudi pomensko ustrezno, je kalkiranje nedvomno najenostavnejši način iskanja slovenskega poimenovanja. Gorjanc (1996: 252) opozarja, da je prav pri kalkiranju metaforičnega poimenovanja večkrat težko govoriti o jezikovnosistemski neustreznosti, kajti metaforična poimenovalna pot kot vir novih

terminoloških poimenovanj zaradi skupne kulturne izkušnje lahko ustreza tudi jeziku prejemnika.

Kot je opaziti denimo na področju računalništva in informatike, je kalkiranje pogosto uporabljeni način tvorjenja slovenskih terminov, ki mu le redko lahko očitamo, da proizvaja pomensko netransparentne ali celo zgrešene rešitve, kot sta denimo *ubežni znak* za *escape character* ali *lahki odjemalec* za *thin client*. Pri številnih novotvorenkah res težko ločimo med kalki in pravimi novimi poimenovanji, zato postopku kalkiranja v tukajšnjem okviru podeljujemo status povsem legitimnega in izredno produktivnega terminotvornega postopka brez negativnih prizvokov, ki se plavajoče umešča nekam med domačo in tujo izvornost.

4.2 TVORJENJE NOVIH POIMENOVANJ V OKVIRU LASTNEGA JEZIKA

Nova poimenovanja se v slovenščini načeloma lahko tvorijo na tri načine (prim. Logar 2005: 213):

- z derivacijo oziroma tvorjenjem besed po jezikovnosistemskih pravilih (npr. *ogled* – *pred-ogled*),
- s pomenotvorjem oziroma dodajanjem pomena obstoječim besedam (npr. *mreža* kot »priprava za ribolov« in *mreža* kot »več medsebojno povezanih računalnikov«),
- s tvorjenjem stalnih besednih zvez (npr. *haptični* + *vmesnik* – *haptični vmesnik*); tu imamo v okviru terminologije v mislih le tvorjenje terminoloških besednih zvez, ne pa denimo splošnojezikovnih frazeoloških enot, kot je *pljuniti v roke*.

Ob tem velja dodati, da so naštet procesi značilni za slovenščino, v drugih jezikih pa obstajajo tudi drugi produktivni načini tvorjenja poimenovanj. Za angleščino je denimo značilna konverzija oziroma prehod iz ene v drugo besedno vrsto brez obrazila (npr. iz glagola *assist* prehod v samostalniik *assist*, ki na področju košarke pomeni *podajo* ali *asistenca*). Ta besedotvorni postopek je redko zaslediti tudi v slovenščini (npr. *Veliko rdeče*). Poleg tega sta angleščina in nemščina izredno gibki pri tvorjenju zloženkov (*compounds*) iz dveh ali več samostalnikov, ki se lahko pišejo skupaj (*teambuilding*), z vezajem (*ground-breaking*) ali narazen (*information system*). Ta način postaja vse dejavnejši tudi v slovenščini, in sicer tako s pisanjem skupaj kot narazen. Logar (2005: 216) navaja naslednje primere iz Dela: *klima naprava*, *kino spored*, *kava bar*, *sendvič stena*, *punkrok*, *bas kitara*, *evroobmočje*, *jazz klub*, *Mobiračun*, *Avtomagazin* itd.

Ob navajanju možnih načinov za tvorjenje poimenovanj pa se postavlja vprašanje, v kolikšni meri so ti postopki sploh predmet terminološke vede. Besede in besedne zveze se v strokovnih jezikih tvorijo po istih načelih, ki veljajo za jezikovni sistem nasploh, se pravi tudi za nastajanje novih besed in poimenovanj v okviru splošnega jezika. Kageura (2002: 48) zato predlaga, naj se terminološka teorija raje sploh ne ukvarja s potencialnimi načini terminotvorja, ampak naj se posveti predvsem svoji opisovalski vlogi. Terminotvorne procese je tako mogoče opisovati le za posamezno strokovno področje, in sicer na podlagi kvantitativnih podatkov, pridobljenih iz reprezentativnega vzorčnega korpusa izbranega področja. Ker ima vsako področje neko pojmovno strukturo, je edini možni način za opis poimenovalnih postopkov izhajanje iz pojmovnega sistema ter pojmovnih kategorij, opaženih na tem področju. Kageura nadalje predlaga tri stopnje opisovanja pojmovnih vzorcev, ki narekujejo terminotvorne procese na določenem področju:

1. Izdelava pojmovnega sistema in opredelitev pojmovnih kategorij ter značilnosti terminov in njihovih sestavnih delov,
2. opredelitev opaženih razmerij med termini oziroma vzorcev za specifikacijo pojmov,
3. opis prednostnih kombinatoričnih vzorcev z uporabo pojmovnih kategorij, razmerij med termini in vzorcev za specifikacijo pojmov, pri čemer ta opis odraža strukturo pojmovnega sistema izbranega področja.

S tem stališčem se Kageura opira na Sagerjev pogled na terminologijo, po katerem je ukvarjanje s termini na vseh ravneh pravzaprav ukvarjanje z besedami. Poseben status terminološka veda doseže šele v svoji kognitivni razsežnosti, kjer preučuje povezave med koncepti ter možnosti njihove formalizacije.

4.3 PREVAJANJE KOT TERMINOTVORNA DEJAVNOST

V številnih strokah pri nas, ki se vsebinsko naslanjajo na tujejezične vire, tvorjenje slovenske terminologije poteka na pasiven način. To v praksi pomeni, da sta najmočnejša kriterija pri izbiri poimenovanj funkcionalnost in gospodarnost, in dokler s strani uporabnikov terminologije ni podana pobuda po dejavnem iskanju slovenskih izrazov, ostaja prevzemanje prevladujoče terminotvorno načelo. Omenjena pobuda se pogosto zgodi v obliki potrebe po prevodu obsežnejšega ali temeljnega dela v slovenščino, ob tem pa se — v idealnem primeru — vzpostavi tudi dialog med jezikoslovci/prevajalci in strokovnjaki z namenom oblikovanja ustrežnejše, enotnejše in morda tudi bolj slovenske terminologije.

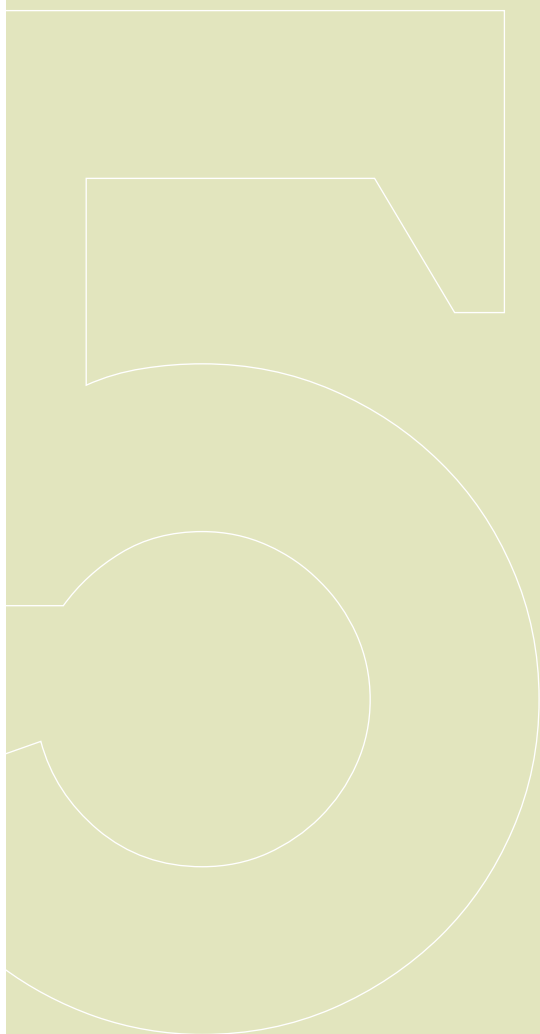
Eno takšnih izkušenj opisuje Kalin Golob (2001), ko se je v sklopu prevajanja besedil za Komunikološko hrestomatijo⁷ porodila ideja o poenotenju slovenskih komunikoloških izrazov. Tako je bila v okviru Komunikološkega društva Slovenije ustanovljena terminološka sekcija, ki se je do izida omenjene publikacije ukvarjala z usklajevanjem, tvorjenjem in predlaganjem slovenskih strokovnih izrazov za to področje.

Prvi korak sekcije je bil nadomeščanje besedne družine latinskega izvora *kommunicirati*, *komuniciranje*, *komunikacija* s slovenskimi *sporočati*, *sporočanje*, *sporočilo*. Latinski izvor je bil ohranjen le pri poimenovanju vede *komunikologija*. Nadalje se za izraza *transmitter* in *receiver* uvajata poimenovanji *oddajni* in *sprejemni pretvornik*, dejavnosti sporočevalca in prejemnika, ki so v tuji literaturi poimenovane kot *šifriranje*, *enkodiranje* in *dekodiranje*, pa so poenotene v *ukodiranje*, *kodiranje* in *razkodiranje*. Prevzete besede angleškega izvora, denimo *feedback*, *input*, *output*, *referent*, *agenda-setting*, *gatekeeper*, so prav tako dobile predlagane slovenske ustreznice: *povratna informacija*, *vnos*, *iznos*, *nanašaneč*, *prednostno tematiziranje*, *odbiratelj*.

Na splošno lahko prispevek komunikološke terminološke sekcije ocenimo kot primer uspešnega posega v strokovnojezikovno rabo, čeprav je bil poseg, kot kaže, bolj usmerjen v slovenjenje kot v poenotenje. Marsikje se namreč za en prevzeti izraz predlaga več slovenskih ustreznic (*telenovela* → *televizijski roman*, *teleroman*). Nekoliko v nasprotju z načelom nedvoumnosti je tudi predlog zamenjave uveljavljenega izraza *mediji z občili*, ki namesto pridevniške rabe predvideva roditeljsko (npr. *medijska vsebina* → *vsebina občil*). Pri takšnih prilastkovnih zvezah, kot opozarja že Kalin Golob (2001: 255), lahko prihaja do pomenskih razlik, zato pridevnik *medijski* ni povsod nadomestljiv z roditeljsko desnoprilastkovno rabo *občil* (npr. *medijsko interpretiranje* → *interpretiranje občil*).

⁷ Splichal, S. (ur.) Komunikološka hrestomatija 1 in 2. Ljubljana: Fakulteta za družbene vede, 2001.

5 Terminografija



Kot je leksikografija dejavnost, ki se ukvarja z zbiranjem in opisovanjem leksemov, se pravi besednih gradnikov jezika na splošno, ter njihovo predstavitev v leksikografskih priročnikih, tako se terminografija ukvarja s pojmi in njihovimi poimenovanji z namenom izdelave terminografskih priročnikov. Bolj nazorno povedano so nameni terminografije naslednji:

- zbiranje in izbiranje terminološkega gradiva
- opisovanje terminologije, ki je trenutno v rabi na določenem področju (deskriptivna terminografija),
- predpisovanje terminologije za določeno področje; svetovanje prednostnih terminov (preskriptivna ali normativna terminografija),
- predstavitev zbranih in obdelanih terminov v terminološki zbirki

Končni cilj vsake terminografske dejavnosti je torej izdelava terminološke zbirke, pa najsi gre za obsežen terminološki slovar, ki pokriva izbrano stroko v celoti, ali za osebni glosarček, ki denimo nastane pri prevajanju določenega projekta.

5.1 TERMINOGRAFSKE DEJAVNOSTI NA SLOVENSKEM

Slovenske eno- in večjezične terminološke zbirke nastajajo na najrazličnejše načine, saj potreba po urejanju in opisovanju strokovnega izrazja včasih izhaja iz same stroke in se tako slovaropisja lotijo kar strokovnjaki, včasih — kot pri Evrotermu — vzgib izhaja iz prevajalskih dejavnosti, še najbolj sistematsko pa se z izdelavo terminoloških slovarjev ukvarjajo sodelavci Sekcije za terminološke slovarje na Inštitutu za slovenski jezik Frana Ramovša. V zadnjih letih je tam nastalo osem obsežnih terminografskih del, in sicer:

Slovenski tehniški slovar (1. snopič A-B) 2007

Geološki terminološki slovar 2006

Geografski terminološki slovar 2005

Gemološki terminološki slovar 2005

Planinski terminološki slovar 2002

Pravni terminološki slovar 1999

Gledališki terminološki slovar 2007

Čebelarski terminološki slovar 2008

Na isti ustanovi so v času pisanja v delu še Botanični terminološki slovar, Farmacevtski terminološki slovar, Slovar elektronike, elektrotehnike in informatike, Slovar kemijske tehnologije, Sodobni pravni slovar, Pravnozgodovinski slovar, Smučarski

terminološki slovar, Terminološki slovar urejanja prostora, Umetnostnozgodovinski terminološki slovar, Slovar evropske zakonodaje in Veterinarski terminološki slovar.

Gantar (2004: 171) opisuje način priprave terminoloških slovarjev na ISJ FR in pri tem opozarja na nekatere pomanjkljivosti. Čeprav sodelavci Inštituta izkazujejo visoko usposobljenost in imajo dolgoletne delovne izkušnje, nabor terminoloških iztočnic za vključitev v slovarje ostaja odvisen od mnenj sodelujočih strokovnjakov, saj sistematskega načina za kvantitativno zbiranje in vrednotenje gradiva ni.

Terminološki slovarji se tako oblikujejo v večletnem obdobju na terminoloških sejah. [...] Glavna pomanjkljivost ostaja v tem, da slovarski opis ne nastaja na podlagi analize sodobnega gradiva, temveč na podlagi jezikovne intuicije skupine avtorjev. [...] Konkretnih aktivnosti v zvezi z izgradnjo terminološkega korpusa glede na trenutne aktualne potrebe za zdaj ni.

(Gantar 2004: 171)

Med cilji, ki bi jih slovenski nacionalni terminološki inštitut moral izpolnjevati, Gantar (2004: 172) omenja opredelitev prednostnih področij, ki naj bi se jim terminografija v prihodnjih letih še posebej posvetila. Ta prednostna področja je mogoče razdeliti na tri skupine, in sicer:

- a) Stroke, za katere je značilen hiter razvoj tehnologije, s tem pa tudi dinamika strokovnih poimenovanj (računalništvo, fizika, kemija, genetika, vojaška industrija),
- b) stroke, ki so bodisi z izrabo novih tehnologij in dosežkov (telekomunikacije, računalništvo, elektrotehnika) bodisi zaradi globalnih dogodkov postale zanimive za širšo javnost (ekologija, alternativna medicina, meteorologija),
- c) stroke, ki doživljajo pospešen razvoj pri nas zaradi nedavnih družbenopolitičnih sprememb in vključevanj v evropske integracijske procese (pravo, ekonomija, borzništvo, menedžment, trženje).

Poleg ustanov, za katere je terminografija eno od osnovnih poslanstev, pa različni glosarji nastajajo tudi v drugačnih okoljih, na eni strani predvsem v gospodarskih družbah oziroma podjetjih, na drugi v akademskih in izobraževalnih okoljih in slednjič pod peresom prevajalcev. Pogled na spletišče Prosto dostopnih slovarjev na internetu⁸ nam naglo ponudi številne primere za vse te skupine:

- s področja bančništva ponujajo glosarčke Abanka, NKBM, SKB, Raffaisen Krekova banka, s področja borzništva pa Ljubljanska borza in Poteza, nadalje objavlja

⁸ <http://evroterm.gov.si/slovar/slovar.html>

Probanka pojasnila v zvezi z vrednostnimi papirji, nenazadnje pa tudi Microsoft na svojem strežniku objavlja prevode svojih izdelkov v 43 jezikih sveta,

- med izobraževalnimi in akademskimi ustanovami, ki objavljajo svoje terminografske dosežke oziroma izdelke svojih študentov, je vodilna Filozofska fakulteta Univerze v Ljubljani z glosarčki astronomije, ekonomije, elektronskega poslovanja itd., poleg nje pa so dejavne še druge članice UL, denimo Biotehniška fakulteta, kjer je nastal obsežen večjezični živalski geslovnik in Agrotezaver, ali Fakulteta za gradbeništvo in geodezijo z Geodetskim slovarjem in tezavrom,
- pomembnejše terminografsko delo, ki je nastalo v prevajalskem okolju, je nedvomno Evroterm, sicer pa glosarčke javno objavljajo tudi prevajalske agencije (BE-Consult), prevajalci, ki delujejo v okviru javnih ustanov (Statistični urad RS) in samostojni prevajalci (Katja Benevol Gabrijelčič, Nada Vukadinovič).

Poseben primer dobre terminografske prakse je Islovar,⁹ terminološki slovar informatike, ki nastaja pod okriljem Slovenskega društva informatika. Slovar zajema informacijsko izrazje, to je temeljno izrazje informatike, informacijske tehnologije in telekomunikacij, pa tudi posebnih področij, kot so baze podatkov, uporabniški vmesniki, poslovna informatika, objektna tehnologija, umetno zaznavanje in sociološki vidiki. Besed splošnega pomena Islovar ne vsebuje.

Islovar nastaja sproti, neposredno na spletu, vanj pa lahko prispeva vsakdo po predhodni prijavi. Vnešene izraze pregleduje in dopolnjuje uredniški odbor, zato so izrazi opremljeni z oznakami ureditve (predlog, pregledano, strokovno pregledano in urejeno). Urejeni sestavki vsebujejo razlage in kvalifikatorje.

Islovar je bil prvi spletni terminografski projekt, ki je že leta 2001 na način wikipedije pričel zbirati izrazje in znanje s področja informatike in računalništva in se tako odzval na problematiko neažurnosti klasičnih (tiskanih) terminografskih priročnikov, še posebej na področjih naglega tehnološkega razvoja. Čeprav je za vzdrževanje takega slovarja potrebno precej truda in dela, bi si sorodnih slovarskih projektov gotovo želeli še več.

Vsem doslej navedenim terminografskim praksam, pa tudi večini avtorjev terminografskih del, je skupna želja po poenotenju izrazja na določenem strokovnem področju in po razširjanju zbranih informacij širši javnosti. Da bi doslej razpršeno terminografsko dejavnost pri nas vsaj v določeni meri poenotili in hkrati ponudili tehnološko podporo terminološkemu delu, je na Filozofski fakulteti Univerze v Ljubljani v teku raziskovalni projekt Slovenski terminološki portal. Cilji projekta so

⁹ <http://www.islovar.org>

opredeliti osnovna terminološka in terminografska načela v skladu z mednarodnimi standardi, zbirati obstoječe terminografske zbirke in jih dati na razpolago v enovitem zapisu TBX, ponujati pretvorbo iz najbolj ustaljenih zapisov terminoloških podatkov v standard TBX in ponuditi spletni dostop do orodja za samodejno luščenje izrazja iz slovenskih strokovnih besedil.

5.2 STRUKTURA TERMINOGRAFSKEGA VNOSA

Obstoječi terminografski priročniki v tiskani obliki, ki jih zasledimo v slovenskem prostoru, so glede strukture gesel in obširnosti opisov tako raznoliki, da bi jih težko celovito predstavili. Oznaka terminološki slovar se v najširšem smislu uporablja za zbirke strokovnih izrazov zelo različnih obsegov, ki so po opremljenosti iztočnic lahko povsem rudimentarne in vključujejo zgolj tujejezično ustreznico, lahko pa v opis dodajajo slovnične podatke, sorodne izraze, definicijo ali razlago, frazeološke podatke, vire, primere rabe, navedbo podpodročja, slikovne prikaze in drugo.

Za ločevanje med različnimi podatkovnimi kategorijami se pri tiskanih priročnikih večinoma uporabljajo tipografske konvencije, ki so v uvodnem delu predstavljene in razložene. V praksi se pokaže, da pri mnogih slovarjih raba teh konvencij ni dosledna in se denimo s poševnim tiskom piše včasih razlaga, včasih pa sopomenka, poleg tega pa je pri večjem številu podatkovnih kategorij tak vnos neizbežno nepregleden, kot ponazarja spodnji primer.

bail *varščina; izpustitev obdolženca na prostost proti varščini (varščina služi kot poroštvo, da se bo obdolženec ob zahtevanem času pojavil na sodišču. Je ukrep za zagotovitev obdolženčeve navzočnosti v kazenskem postopku namesto pripora in predstavlja denar ali premoženje, ki ga obdolženec ali oseba, ki zanj jamči, položi na sodišču kot varščino.); porok (oseba, ki jamči, da se bo obdolženec ob zahtevanem času pojavil na sodišču)*

Vir: Sket: Angleško-slovenski in slovensko-angleški terminološki slovar kriminologije in kazensko-pravnih znanosti

Ker so tiskani terminološki slovarji tipično urejeni abecedno, je v njih težko ustrezno predstaviti razmerja med pojmi, še težje pa je ob abecedni ureditvi udejanjati pojmovni pristop, ki narekuje en vnos za vsak pojem oziroma pomen.

V nadaljevanju pričujočega razdelka tako govorimo o terminoloških bazah, ki zaradi elektronske oblike in sodobnih programskih orodij, s katerimi jih gradimo, avtorju in/ali uporabniku omogočajo lastno zasnovo strukture terminološkega vnosa.

Podatkovne kategorije, s katerimi opisujemo terminološke podatke, se delijo na tri ravni (ISO:12620 Data Categories):

- a) tiste, ki se nanašajo na pojem: **definicija, razlaga, slika ali grafični prikaz, področna klasifikacija,**
- b) tiste, ki se nanašajo na termin: **slovnične kategorije (spol, število, oblikoslovne posebnosti), kolokacije, primer rabe,**
- c) tiste, ki podajajo administrativne podatke za lažje upravljanje baze: **številka ali ID vnosa, datum nastanka vnosa, datum zadnje spremembe, avtor vnosa, avtor spremembe itd.**

Pri uporabi določenih programskih orodij, denimo programa MultiTerm, pa tudi pri zapisovanju terminoloških baz v skladu s standardom TBX se pojavi zadrega, kajti čeprav so določeni podatki sicer vezani na pojem in tako teoretično jezikovno neodvisni, so vendarle zapisani v določenem jeziku. Razlaga, ki sicer sodi na raven pojma, se tako glede jezika, v katerem je oblikovana, pridružuje skupini terminov tistega jezika. Iz tega razloga se v zgornjo shemo uvede še raven jezika, ki je vmesna med pojmovno in terminološko. Tabela 4 navaja izbrane podatkovne kategorije po posameznih ravneh, povzete in prirejene po standardu TBX.

Tabela 4: Podatkovne kategorije po ravneh

Na ravni termina
Tip (polni termin, krajšava, simbol, formula)
Opomba
Besedna vrsta (sam., gl., prid., prisl.)
Spol (m, ž, s, drugo)
Število (ednina, dvojina, množina, drugo)
Register (nevtralno, tehnično, specifično za vir, pogovorno, slengovsko, vulgarno)
Pogostost (redko)
Časovna oznaka (zastarelo, novo)
Zaščiteno (blagovna znamka, zaščiteno ime)
Izgovarjava
Jezikovnonačrtovalni kvalifikator (priporočeno, nestandardizirano, predlog)
Stopnja ustrezanja (1 2 3 4 5 6 7 8 9 10)
Primer
Kolokacije
Vir
Tip vira (vzporedno besedilo, dodatno gradivo, drugo)

Na ravni jezika

Definicija

Vir

Tip vira

Razlaga

Sorodni pojmi

Opomba

Na ravni pojma

Področje

Klasifikacija

Slika

Zvočni posnetek

Video posnetek

Tabela

Vir

Tip vira (vzporedno besedilo, dodatno gradivo, drugo)

Opomba

Nabor navedenih podatkovnih kategorij je obsežen, a ne popoln. Prav vsak terminografski projekt ima svoje značilnosti, zato je nemogoče podajati splošna priporočila za strukturo terminoloških baz. Uporaba zgornjih kategorij in napotkov, ki so tudi del Slovenskega terminološkega portala, pa morda pomeni korak k boljši izmenljivosti in združljivosti različnih virov.

Pri snovanju strukture terminološkega vnosa in še posebej pri vnosu podatkov v bazo pa je koristno upoštevati še tri pomembna načela (povzeto po eCoLoTrain: Terminology Management II):

1. **Elementarnost**, s čimer poimenujemo načelo, da se v vsako podatkovno kategorijo zapiše le en podatkovni element. Če je polje za slovenski izraz denimo izpolnjeno z *jezikovne tehnologije (JT)*, smo vanj zapisali dva elementa, in sicer termin *jezikovne tehnologije* in njegov akronim na ravni sinonima *JT*. Pravilno bi bilo ta dva podatka vnesti v ločeni podatkovni polji.
2. **Granularnost**, ki se nanaša na načelo, da naj bodo podatki, vsebovani v enem podatkovnem polju, istega tipa. Če baza vsebuje polje Slovnčni podatki, kamor zapišemo *m*, *pl* kot oznako za moški spol in množino, smo v eni podatkovni kategoriji pomešali dve različni slovnčni informaciji. Bolje je torej definirati polji Spol in Število.

3. **Soodvisnost kategorij**, s čimer izražamo razmerja med posameznimi podatkovnimi polji. Kategorija Vir lahko vsebuje podatke o viru termina, razlage, primera rabe ali celotnega vnosa. Za njeno pravilno tolmačenje mora biti torej iz strukture terminološkega vnosa jasno razvidno, h kateri nadrejeni kategoriji sodi.

Pri gradnji terminološke baze je prvi korak navadno izdelava geslovnika, se pravi nabor in izbor gesel, ki bodo predstavljena v bazi. Podrobneje o metodah, ki nam pomagajo pri izdelavi geslovnika, govorimo v razdelku 7.2, tu pa naštujemo le nekaj načel, ki jih moramo upoštevati ne glede na izbrano metodologijo.

- a) **Ciljni uporabniki**. Kot smo poudarili že na več mestih te knjige, se pojmovanje terminov izredno razlikuje glede na uporabniški vidik. Podrobna opredelitev ciljnih uporabnikov vpliva na izbor vnosov ter način njihove predstavitve.
- b) **Področje**. Pri izbiri gesel se pogosto znajdemo pred težavno odločitvijo, ali določeni termin še sodi v izbrano področje ali ne. Ob reševanju te dileme se je koristno vprašati, ali poznamo »matično« področje spornega izraza in ali je prekrivanje med našim izbranim področjem ter matičnim področjem spornega izraza širše, tj. sistemsko, ali gre za tematsko ali kontekstno pogojeno povezavo gostujočega izraza z našim področjem.
- c) **Uravnotežena razvejanost**. Tu gre predvsem za spoštovanje pojmovnega načela tudi v tem smislu, da baza za določeni pojem vsebuje njegovo celotno pojmovno polje, ne pa da v geslovník vključimo le eno od podpomerk, na ostale pa pozabimo.

5.3 DEFINICIJE IN RAZLAGE POJMOV

Predstavitve pojma v terminološki bazi ali slovarju pogosto vsebuje tudi opredelitev njegovega pomenskega polja. Temu opisu pravimo definicija, kadar je oblikovan v skladu z v nadaljevanju predstavljenimi načeli, sicer pa razlaga. Razlaga je torej nekoliko širši in ohlapnejši izraz kot definicija, čeprav se v določenih terminografskih okoljih izraz razlaga uporablja kot slovenska ustreznica definicije.

Standard ISO 1087 definicijo opredeljuje takole:

A definition is a statement which describes a concept and permits its differentiation from other concepts within the system of concepts.

(ISO 1087)

Osnovno in že tradicionalno načelo pri pisanju definicij je, da se pojem opredeli z navedbo najbližjega nadrejenega pojma in značilnosti, ki pojem razlikujejo od nadrejenega in istorednih pojmov v sistemu (*per genus et differentiam*). Tako se z ustrezno oblikovano definicijo obenem vzpostavlja pojmovni sistem terminološke zbirke in vnosov ni potrebno posebej opremljati s polji za nadpomenke in sorodne pojme. Primer je iz Pojemovnika jedrske tehnike in varstva pred sevanji:¹⁰

kroglični reaktor - Reaktor, v katerem je del snovi ali vsa snov (npr. gorivo, oplodna snov, moderator) v obliki mirujočega kupa majhnih krogel, ki se dotikajo ena druge.

Izrazi, ki jih uporabimo v definiciji, naj bi bili definirani v okviru iste publikacije oziroma se moramo nanje sklicevati, razen če gre za izraze splošnega jezika. Izogibati se moramo tudi krožnim definicijam, kar pomeni, da pojem A definiramo z izrazom B, slednjega pa spet z A. Iz publikacije (predgovora, navodil za uporabo ipd.) naj bi bilo razvidno, po katerih načelih so definicije oblikovane in komu so namenjene, saj definicija v srednješolskem učbeniku ni enaka definiciji v znanstvenem priročniku. Definicija naj bi predstavljala čimbolj jedrnat opis pojma, kar pa ni nujno za razlago, ki lahko vključuje tudi dodatne podatke o pojmu in pojasnjuje njegov pomen v širšem smislu.

Poznamo dva tipa definicij, ki lahko nastopata tudi skupaj:

- intenzionalna — označi lastnosti pojma, ki ga želimo definirati, se pravi intenzijo,
- ekstenzionalna — definira pojem tako, da našteje vse pripadnike pojma na isti posplošitveni ravni, ali vse predmete, ki jih razumemo pod tem pojmom.

Spodaj navajamo nekaj primerov. Razlaga za mesojedo rastlino je primer intenzionalno-ekstenzionalne definicije.

grebénska opást *opastí* ž opast, ki s privetrne strani grebena sega nad odvetrno strmino
Vir: Humar, M. (in drugi uredniki): Planinski terminološki slovar, Ljubljana, ZRC SAZU, 2002.

mesojéda rastlína -e -e m nav. mn. ekol. *rastline, ki z različno oblikovanimi lovilnimi napravami lovijo majhne živali, zlasti žuželke, redko majhne sesalce, ptiče, in jih prebavljajo, da dobijo predvsem manjkajoči fosfor in dušik, npr. vrčnica (Nepenthes), mešinka (Utricularia), rosika (Drosera), muholovka (Dionea) **S:** insektivorna rastlina, karnivorna rastlina, žuškojéda rastlína*

ang.: carnivorous plant n.

nem.: carnivore Pflanze f., tierfangende Pflanze f.

lat.: plánta carnívora f.

Vir: Botanični terminološki slovar (v nastajanju, ZRC SAZU).

¹⁰ <http://www.drustvo-js.si/pojmovnik/index.php>

hard radiation *trdo sevanje*

reaktorska fizika

Ionizirajoče sevanje z visoko energijo, ki ima visoko sposobnost prodiranja skozi snov.

Vir: Pojmovnik jedrske tehnike in varstva pred sevanji

copy *besedilo oglasa*

Besedilo tiskanega ali predvajanega oglasa skupaj z aslvi in podnaslovi, skicami, slikami ipd. z določeno sporočilno učinkovitostjo.

Vir: Potočnik, V. in Umek, A. (2004) Terminološki slovar trženja. Založba GV, Ljubljana.

5.4 TERMINOGRAFIJA ZA PREVAJALSKE NAMENE

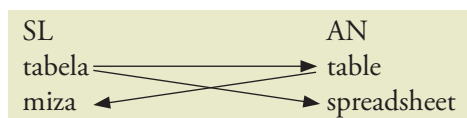
Pri prevajanju specializiranih besedil prevajalci posegajo po vseh razpoložljivih virih, zato je vsak terminološki slovar, če je le dovolj sodoben in kredibilen, koristen pripomoček za prevajalca. Po vsej verjetnosti se tudi vsak avtor večjezičnega terminološkega priročnika v kakršni koli obliki počuti zaslužnega za poslej lažje življenje prevajalcev besedil tega področja.

V resnici pa je le malo terminoloških priročnikov v tiskani obliki — in nekaj več v elektronski — zasnovanih v skladu s specifičnimi prevajalskimi potrebami, saj se o njih razmeroma malo govori in piše. Prevajalec se kot uporabnik terminološkega priročnika razlikuje od strokovnjaka po nekaterih bistvenih značilnostih:

- a) **Poznavanje področja.** Čeprav se mnogi prevajalci specializirajo za določeno področje in ga po dolgoletni prevajalski praksi tudi že podrobno poznajo, za večino prevajalcev strokovnih besedil vendarle velja, da njihova raven poznavanja stroke ne dosega tiste, ki jo srečamo pri »pravih« strokovnjakih, se pravi tistih, ki se z določenim področjem poklicno ukvarjajo in se na tem področju tudi izobražujejo.
- b) **Jezikovno in jezikoslovno znanje.** Prevajalec kot uporabnik terminološkega priročnika načeloma aktivno obvlada vsaj dva od vsebovanih jezikov, zato je sposoben uporabiti tudi razlage in druge podatke, zapisane v tujem jeziku. Po drugi strani potrebuje manj jezikoslovnih podatkov o terminih (kot so izgovorjava, sklanjatev in spol v domačem jeziku).
- c) **Način uporabe terminološkega priročnika.** Prevajalec ga uporablja z namenom, da oblikuje besedilo v ciljnem jeziku, ki bo pomensko in funkcionalno čim bolj ustrezalo danemu izvirniku. Strokovnjaki sicer tudi pogosto posegajo po večjezičnih terminoloških virih takrat, ko tvorijo besedila v tujem jeziku, vendar je med prevajanjem in tvorjenjem besedila brez predloge velika

razlika. Pri slednjem je v primeru terminološke zagate ali vrzeli v jezikovnem znanju besedilo mogoče prilagoditi in željeno ubesediti na drugačen način, prevajanje strokovnega besedila pa je vezano na izvirnik in ne dopušča večjih izpustov.

Iz teh razlik izhajajo nekateri napotki za snovanje terminoloških priročnikov, da bodo ti čim bolj pisani na kožo prevajalcem. Predvsem naj bi bili večjezični terminološki priročniki, in tu mislimo v prvi vrsti na terminološke baze v elektronski obliki, zasnovani po onomaziološkem oziroma pojmovnem načelu. To namreč ne le omogoča lažjo pojmovno strukturiranost zbirke, ampak ima pri večjezičnih bazah še bolj primarno funkcijo — omogoča obrnljivost glosarja oziroma baze. Spodnji primer nam kaže prepletanje pomenov, ki nastane kot rezultat semaziološkega pristopa v terminologiji. Tako zasnovani glosar je lahko samo enosmeren, njegova izdelava v obrnjeni smeri je lahko zahteven projekt.



Če je v terminološko bazo vključenih več jezikov, in nekatere obstoječe zbirke vsebujejo denimo vse jezike EU, v času pisanja triindvajset, je pojmovni pristop edini možni pristop, saj se sicer zapletemo v nepregledno mrežo večpomenskih izrazov ali — na drugi strani — leksikalnih praznin.

Nadalje naj bi terminološki priročniki za prevajalske namene vsebovali razlage pojmov, ki omogočajo spoznavanje področja in hkrati razkrivajo osnovna medpojmovna razmerja. Razlaga navadno navaja najbližjo nadpomenko pojma, pogosto pa omenja tudi sorodne pojme, kar prevajalcu olajšuje ustvarjanje miselne predstave o prevajanem besedilu.

Posebej pomembna lastnost prevajalcem namenjenih priročnikov pa je dobra predstavitev terminološke frazeologije, se pravi morebitnih kolokacij, skladenjskih posebnosti idr. Prav v tem segmentu so obstoječi terminološki priročniki najrevnejši, saj le redki vsebujejo posebne podatkovne kategorije za frazeološke podatke, kot so kolokacije in primeri rabe. Za pridobivanje teh podatkov je smiselno uporabiti korpus strokovnih besedil, pri večjezičnem slovarju pa je seveda idealno, če imamo na razpolago vzporedni korpus področja.

Za prevajalce je posebnega pomena tudi dokumentiranost terminoloških podatkov, uvrščenih v bazo, da lahko bolj presojajo o ustreznosti in zanesljivosti predlaganega

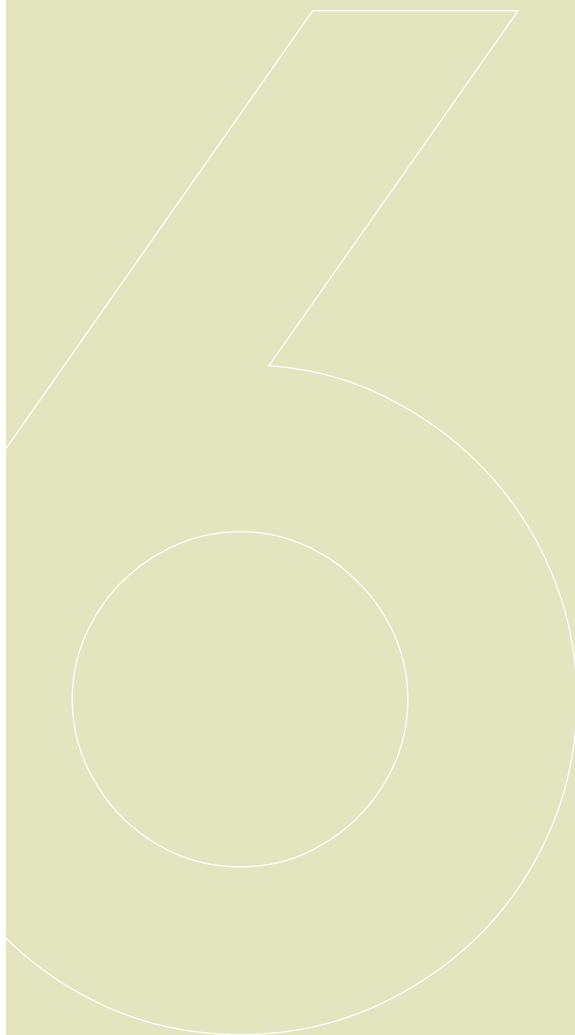
izraza. Če slovar denimo vsebuje več poimenovanj za določeni pojem, je za prevajalca koristno vedeti, od kod izvirajo našeta poimenovanja in kje so bila že uporabljena. Na tak način je zasnovana tudi baza Evroterm, ki prek podatkovnih polj TermRef, TermSource, DefRef in Context dokumentira izvor izraza, s poljema Note in Reliability pa podaja še podrobnejše informacije o rabi in zanesljivosti (Slika 8).¹¹

Slika 8: Dokumentiranje podatkovnih polj v Evrotermu

Subj	administracija
Subject	administration AD
Note	prevod iz slovenščine
ID	20569
SL	<u>uradna oseba</u>
	TermRef Ustava Republike Slovenije, 36. člen
SL	<u>uradnik</u>
	Reliability 5
	Definition Uradniki so javni uslužbenci, ki v organih opravljajo javne naloge. Javne naloge v organih so naloge, ki so neposredno povezane z izvrševanjem oblasti ali z varstvom javnega interesa.
	DefRef Zakon o javnih uslužbencih, 23. člen
EN	<u>official</u> 
	TermRef Constitution of the Republic of Slovenia
	TermRef1 prevod Zakona o javnih uslužbencih - Civil Servants Act

¹¹ Tudi Evroterm po svoji strukturi in vsebini ni zgledna terminološka baza za prevajalske namene, česar se zavedajo tudi njegovi uredniki. Zaradi izredne obsežnosti in različnih virov vnašanja baza še vedno vsebuje neprečiščene in podvojene vnose.

6 Standardizacija terminoloških virov



Standardizacija je dejavnost vzpostavljanja usklajenih pravil in določil za ponavljajočo se uporabo, da se doseže optimalna stopnja urejenosti na nekem področju. Njeni temeljni cilji so že od zgodovinskih začetkov usmerjeni v preprečevanje ovir v trgovanju ter izboljšavo pogojev za sodelovanje in izmenjavo. Standardizacija se zato nanaša predvsem na izdelke in storitve, proizvodne metode in postopke, merske enote, varnost ljudi in blaga ter terminologijo in simbole, o čemer podrobneje govorimo v nadaljevanju.

Na splošno velja, da je za specializirane oblike komunikacije, za razliko od splošne, zahtevana višja stopnja natančnosti. V splošnem jeziku so mnoge besede večpomenske, obenem pa lahko obstaja za določeni pojem več izrazov, ki so med sabo sopomenski. Ker je v specializiranem jeziku taka situacija nezaželena, saj slabo vpliva na jasnost in učinkovitost komunikacije, je osnovni namen standardizacije terminologije uveljavljanje načela en pomen — en izraz.

S tega vidika ima izraz standardizacija terminologije dva pomena, saj v širšem pomenu vključuje vso dokumentirano in javno terminografsko dejavnost. Povedano drugače, izid terminološkega slovarja ali terminološke baze na spletu je nedvomno standardizacijski akt v smislu, da bistveno pripomore k poenotenju izrazja določenega področja. V ožjem pomenu se standardizacija nanaša na institucionalno normiranje izrazja, ki ga izvajajo mednarodni in nacionalni organi za standardizacijo, konkretno tehnični odbor ISO/TC 37 in njegovi štirje pododbori, na nacionalni ravni pa Slovenski inštitut za standardizacijo (SIST).

6.1 MEDNARODNI TERMINOLOŠKI STANDARDI

V mednarodnem prostoru je tehnični odbor Izrazoslovje v okviru Mednarodne zveze standardizacijskih združenj (ISA) že leta 1947 začel uresničevati standardizacijo terminoloških načel in metod za terminološko delo, pripravo slovarjev in računalniško podprto slovaristiko. S pripravo metodoloških terminoloških standardov je po ustanovitvi Mednarodne organizacije za standardizacijo (ISO) leta 1952 njegovo delo nadaljeval tehnični odbor ISO/TC 37 Izrazoslovje — načela in koordinacija. ISO/TC 37 ima danes naziv Terminologija, jezikovni in drugi vsebinski viri, njegov cilj pa je formuliran kot standardizacija načel, metod in aplikacij v zvezi s terminologijo in drugimi jezikovnimi in vsebinskimi viri v okviru večjezične komunikacije in kulturne raznolikosti. Razdeljen je na štiri pododbore, ki pokrivajo različne s terminologijo povezane veje.

- **ISO/TC 37/SC1: Načela in metode terminologije**

Cilj: Standardizacija temeljnih načel in metod za razvoj znanstvenih in tehničnih terminoloških zbirk in drugih jezikovnih virov.

Najpomembnejši objavljeni standardi:

ISO 704:2000 Terminology work – Principles and methods

ISO 860:1996 Terminology work – Harmonization of concepts and terms

ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application

- **ISO/TC 37/SC2: Terminografija in leksikografija**

Cilj: Standardizacija terminoloških in leksikografskih delovnih metod, postopkov in kodnih sistemov ter upravljanje kulturne raznolikosti

Najpomembnejši objavljeni standardi:

ISO 639-1:2002 Codes for the representation of names of languages – Part 1: Alpha-2 code

ISO 639-2:1998 Codes for the representation of names of languages – Part 2: Alpha-3 code

ISO 1951:1997 Lexicographical symbols and typographical conventions for use in terminography

ISO 10241:1992 International terminology standards -- Preparation and layout

ISO 12199:2000 Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet

ISO 12616:2002 Translation-oriented terminography

ISO 15188:2001 Project management guidelines for terminology standardization

- **ISO/TC37/SC3:Sistemizaupravljanjeterminologijeinpovezljivostvsebin**

Cilj: Standardizacija načel in zahtev za semantično povezljivost, terminologijo in sisteme za upravljanje vsebine ter urejanje znanja

Najpomembnejši objavljeni standardi:

ISO 1087-2:2000 Terminology work – Vocabulary – Part 2: Computer applications

ISO 6156:1987 Magnetic tape exchange format for terminological/ lexicographical records (MATER) – umaknjen

ISO 12200:1999 Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange

ISO 12620:1999 Computer applications in terminology – Data categories

ISO 16642:2003 Computer applications in terminology – Terminological markup framework

- **ISO/TC 37/SC4: Upravljanje jezikovnih virov**

Cilj: Standardizacija specifikacij za računalniško-podprto upravljanje z jezikovnimi viri

Najpomembnejši objavljeni standardi:

Ker gre za najmlajši pododbor, objavljenih standardov še ni.

Standardi v pripravi:

ISO/AWI 21829	Terminology for language resources
ISO/CD 24610-1	Language resource management – Feature structures – Part 1: Feature structure representation
ISO/WD 24611	Language resource management – Morphosyntactic annotation framework
ISO/WD 24612	Language Resource Management – Linguistic Annotation Framework
ISO/WD 24613	Language resource management – Lexical markup framework
ISO/AWI 24614-1	Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 1: General principles and methods
ISO/AWI 24614-2	Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 2: Word segmentation for Chinese, Japanese and Korean
ISO/NP 24614-3	Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 3: Word segmentation for other languages

Posebej pomembno področje standardizacije je zapis terminoloških baz. Že pred več kot dvema desetletjema se je zaradi obstoja številnih komercialnih in prosto dostopnih orodij za upravljanje terminologije pokazala potreba po boljši izmenljivosti in povezljivosti terminoloških baz, zato je bil že leta 1987 sprejet standard MATER, ki pa se je nanašal na poenotenje zapisov na magnetnih trakovih in je bil že ob objavi zastarel. Dobrih deset let kasneje je vrzel skušal premostiti standard MARTIF (Računalniško berljiv zapis za izmenjavo terminoloških podatkov). Medtem je eXtensible Markup Language ali XML¹² zrasel v prevladujoči zapis ne le za jezikovne podatke, ampak tudi za številne druge vrste strukturiranih podatkovnih baz, in na njem danes temeljijo tudi priporočila za zapis besedilnih podatkov TEI (Text Encoding Initiative)¹³.

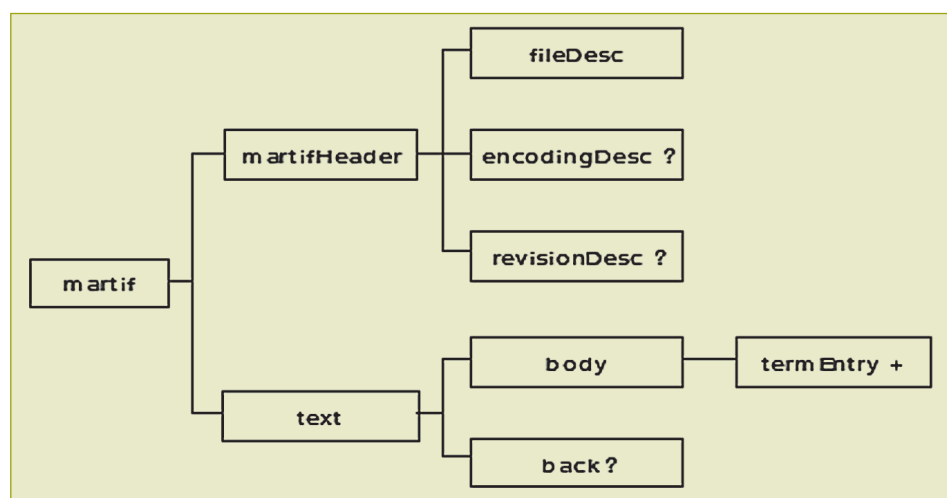
¹² <http://www.w3.org/XML/>

¹³ <http://www.tei-c.org/>

Maja 2002 je Organizacija za standardizacijo lokalizacijske industrije (LISA) izdala prvo različico novega predlaganega standarda za izmenjavo terminoloških podatkov **Term Base eXchange** (TBX), ki je trenutno v postopku standardizacije kot ISO DIS 30042 (Melby 2003). Osnutek je dostopen tudi na spletnih straneh Lise.¹⁴ Ker gre za standard, ki se je še pred uradnim sprejetjem uveljavil v številnih industrijskih aplikacijah, ga v nadaljevanju podrobneje predstavljamo.

TBX ohranja osnovno strukturo terminološkega vnosa predhodnega standarda MARTIF, zato v opredelitvi strukture najdemo krovni element `martif` in glavo `martifHeader` (Slika 9), v kateri so upravni podatki o terminološki bazi. Terminološki podatki sodijo v element `body`.

Slika 9: Krovna struktura baze v TBX



Vsak pojem je predstavljen v elementu `termEntry`, ki je nato sestavljen iz opisnih polj `descrip` in polj za posamezne jezike `langSet`. Slednji naprej vsebuje bodisi element `tig` bodisi `ntig`, ki je sestavljen iz samega termina `term`, opomb `termNote`, opisov `descrip` in drugih možnih elementov. Termin je vsebovan v elementu `term`.

¹⁴ <http://www.lisa.org/Term-Base-eXchange.32.0.html>

Slika 10: Primer terminološkega vnosa v TBX

```
<?xml version='1.0'?> <!DOCTYPE martif SYSTEM "TBXcoreStructV02.dtd">
<martif type="TBX" xml:lang="en">
  <martifHeader>
    <fileDesc>
      <sourceDesc>
        <p>from an Oracle corporation termBase</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <p type="DCSName">TBXXCSV02.XCS</p>
    </encodingDesc>
  </martifHeader>
  <text>
    <body>
      <termEntry id="eid-Oracle-67">
        <descrip type="subjectField">manufacturing</descrip>
        <descrip type="definition">A value between 0 and 1 used in ...</descrip>
        <langSet xml:lang="en">
          <tig>
            <term id="tid-Oracle-67-en1">alpha smoothing factor</term>
            <termNote type="termType">fullForm</termNote>
          </tig>
        </langSet>
        <langSet xml:lang="hu">
          <tig>
            <term id="tid-Oracle-67-hu1">Alfa simítási tényező</term>
          </tig>
        </langSet>
      </termEntry>
    </body>
  </text>
</martif>
```

TBX je zelo fleksibilen glede tipov podatkovnih kategorij, ki jih vnašamo v bazo, in opredeljuje 17 podatkovnih kategorij na ravni pojma (npr. *definition*, *explanation*, *figure*), 17 medpojmovnih razmerij (npr. *broaderConcept*, *Generic*, *subordinateConceptPartitive*), 19 kategorij, ki natančneje opredeljujejo termin (npr. *PartOfSpeech*, *animacy*, *grammaticalGender*), 7 kategorij o tipu termina (npr. *abbreviatedFormFor*, *shortFormFor*) in 17 kategorij za upravne podatke o vnosih (npr. *administrativeStatus*, *originatingPerson*, *originatingInstitution*).

6.2 STANDARDIZACIJA V SLOVENIJI

S standardizacijo se v Sloveniji ukvarja Slovenski inštitut za standardizacijo (SIST), ki opravlja naloge v skladu z Zakonom o standardizaciji. V to delo sta vključena

tudi terminološko delo in standardizacija izrazja s strokovnih področij dejavnosti, ki sodijo v pristojnost SIST.

Terminološke dejavnosti se v SIST izvajajo po naslednjem vrstnem redu (Hočevar in Kiralj 2004: 203):

- privzem in prevod mednarodnih terminoloških pravil (ISO TC 37),
- prevodi mednarodnih in evropskih terminoloških standardov (npr. SIST EN 45020),
- prevodi standardizacijskih dokumentov, vodil (standardizacija kot stroka),
- prevodi naslovov standardov EN,
- prevodi evropskih standardov, na katere se sklicuje zakonodaja,
- prevodi osnovnih oziroma temeljnih mednarodnih in evropskih standardov (npr. skupina standardov s področja vodenja in zagotavljanja kakovosti SIST EN ISO 9000, ugotavljanja skladnosti SIST EN 45000, veličine in enote SIST ISO 31, tehnična dokumentacija),
- prevodi vsebin drugih evropskih in mednarodnih standardov.

Hočevar in Kiralj (2004: 205) poudarjata, da je priprava prevodov terminoloških standardov znotraj posameznih strok nujna, prednostna in stalna naloga. To je podlaga za poznejšo pripravo ustreznih prevodov drugih mednarodnih in evropskih standardov oz. standardizacijskih dokumentov. Pri prevodih terminoloških standardov je treba uporabljati izraze, ki so že ustaljeni, in opuščati izraze, ki se v praksi ne uporabljajo. Glede računalniške podpore terminološkemu delu tudi sodelavca SIST-a ugotavljata, da je najprimernejša programska oprema Trados MultiTerm, ker omogoča lastno določanje strukture baze, podpira XML, deluje v mreži in ga uporablja precej vidnih ustanov v Sloveniji in v tujini.

Če v sklepnem delu tega poglavja spregovorimo še o dostopnosti standardov na nacionalni in mednarodni ravni, je za začetek smiselno omeniti, da so standardi, ki jih izdaja Mednarodna organizacija za standardizacijo (ISO), plačljivi in stanejo od 100 do 300 švicarskih frankov vsak. Javno razširjanje besedil standardov, se pravi kopiranje ali objava na spletu, je strogo prepovedano. V času, ko se dostopnost vseh vrst informacij na spletu skokovito povečuje in so uporabniki navajeni plačevati le za tiste informacijske storitve, ki jih res potrebujejo in jih nikakor ne morejo dobiti zastonj, plačljivost standardov pomeni manjše število uporabnikov, kot če bi bili ti zastonj. Pri dejavnosti, kot je standardizacija, katere namen je vsaj posredno tudi vzgojen v smislu širjenja dobre prakse med uporabniki, je taka situacija zagotovo nekoliko absurda.

Ne dosti drugačno je stališče SIST-a, ki sicer v vseh svojih gradivih o standardizaciji terminologije poudarja dejavno sodelovanje s strokovnjaki, zamolči pa dejstvo, da morajo člani strokovnih odborov SIST za svoje delo SIST-u plačevati letno članarino v nezanemarljivem znesku, in to ne glede na to, ali v tem odboru potekajo kake terminološke dejavnosti ali ne.

Izkušnje z razvojem spleta kažejo, da morajo biti vsaj standardi za zapis računalniških podatkov javni in prosto dostopni, saj sicer nikoli ne dočakajo pravega statusa standarda. Na področju strokovnih jezikov in njihove dejanske rabe pa je vpliv standardizacije v ožjem pomenu zelo majhen.

7 Računalniško podprta terminografija



7.1 GRADNJA SPECIALIZIRANIH KORPUSOV

Internet je danes neizčrpen vir strokovnih besedil, zastopanost posameznih področij pa je vse bolj uravnotežena. Za številne »priložnostne« terminološke potrebe, denimo nekatere tipe prevajalskih projektov, si je tako mogoče zgraditi specializirani korpus iz spletno dostopnih besedil. Nekaj takih korpusov je celo že na razpolago, tudi dvo- in večjezičnih. Novembra 2007 je Generalni direktorat Evropske komisije za prevajanje (DGT) v javno uporabo predal pomnilnike prevodov, ki so nastali ob prevajanju *acquis communautaire*. Gre torej za zbirko vzporednih besedil v 22 jezikih, kar omogoča izdelavo vzporednih korpusov za 462 jezikovnih parov.¹⁵ Seveda pa je to le eden od številnih večjezičnih besedilnih virov, ki jih ponujajo spletišča EU.

Za potrebe podpore terminološkemu delu pa je pri zbiranju in izbiri besedil pomembno upoštevati nekatere kriterije, ki ključno vplivajo na uporabnost korpusa za terminološke namene. V nadaljevanju razpravljamo o nekaterih vidikih tipologije specializiranih korpusov.

7.1.1 Vrste korpusov

Jezikovne korpuse delimo na več vrst glede na različne vidike. Mednje sodijo naslednji pomembni vidiki, ki nam pomagajo opredeliti tip korpusa (prim. Erjavec 1996/97, Gorjanc 2005):

Obseg

Glede na obseg jezikovnih zvrsti, ki naj bi bile v korpusu zaobjete, poznamo **referenčne korpuse** in **specializirane korpuse**. Referenčni korpus je enojezikovna zbirka besedil, ki naj bi predstavljala celovito podobo nekega jezika in tako služila kot izhodišče za temeljne jezikovne raziskave. Specializirani korpus je reprezentativni vzorec jezikovnih zvrsti v okviru določenega strokovno, socialno ali geografsko opredeljenega specialnega jezika.

Jezik

Poznamo **enojezikovne** in **večjezikovne korpuse**. Slednji se delijo na **vzporedne** in **primerljive**, kjer vzporedni korpus vsebuje poravnana besedila v izvirniku in najmanj enem prevodu, primerljivi korpus pa vsebuje besedila, ki so primerljiva glede na žanrsko opredelitev, register, temo ipd.

¹⁵ <http://langtech.jrc.it/DGT-TM.html>

Časovni razpon

Korpus lahko predstavlja jezik določenega obdobja v **zaključeni** besedilni zbirki, lahko pa je zasnovan kot **spremljevalni** korpus, pri katerem se sproti vključujejo nova besedila in izločajo stara. **Diahroni** korpusi predstavljajo prerez skozi zgodovino jezika in zajemajo besedila širšega časovnega razpona.

Medij

Korpus lahko vsebuje **pisna** ali **govorna** besedila. Danes referenčni korpusi večinoma vključujejo tudi transkripcijo govora, sicer pa se govorni korpusi zaradi bistveno drugačne metodologije oblikujejo samostojno (Gorjanc 2005: 8). Govorni korpusi lahko vsebujejo transkripcije branih besedil, spontanega govora in pogovora. V zadnjem času se na področju elektronske komunikacije pojavljajo hibridni mediji, ki v pisni obliki kažejo izrazite značilnosti govornih besedil (npr. SMS sporočila, forumi in blogi).

Označenost

Glede na raven jezikoslovne analize lahko korpuse delimo na **neoznačene**, **obliko-skladenjsko** označene, **polno razčlenjene** (kjer so besedila označena s polno skladenjsko strukturo) in druge, saj se glede na namen korpusa pogosto označujejo tudi druge jezikoslovne prvine.

Med najbolj znane referenčne korpuse sodijo British National Corpus in Bank of English za angleški jezik, kjer je slednji spremljevalnega tipa, češki nacionalni korpus, za slovenski jezik pa Fidaplus.

Med specializiranimi korpusi je prav tako največ angleških, predstavljajo pa širok razpon različnih vrst specialnih jezikov. Tako je na voljo več korpusov učenja jezika, govorni korpusi posameznih področij, na primer transkripcije pogovorov med potovalnimi agencijami in strankami, pogovori med kontrolorji letenja in piloti, strokovni korpusi akademskih člankov in doktorskih disertacij, zbirke poslovnih, pravnih, računalniških besedil in številnih drugih področij.

Nadalje se za sociolingvistične namene zbirajo korpusi posameznih sociolektov, na primer korpus londonskega najstniškega govora COLT in korpusi dialektov in regionalnih jezikovnih variant.

Vzporedni korpus je posebej dragocen jezikovni vir za prevodoslovne in kontrastivne raziskave, obenem pa je to eden zahtevnejših korpusnih tipov, saj je tako z vidika zbiranja besedil kot z vidika poravnave in označevanja vanj potrebno vložiti ogromno dela. Za slovensko-angleški jezikovni par imamo v času pisanja na voljo tri prosto dostopne vzporedne korpuse:

- **IJS-ELAN.** Ta korpus je nastal v okviru mednarodnega EU projekta ELAN (European Language Activity Network) in vsebuje 15 poravnanih slovensko-angleških besedil v skupnem obsegu milijon besed. Korpus je prosto dostopen za raziskovalne namene in za spletno iskanje.¹⁶
- **Evrokorpus.** To je korpus prevodov zakonodaje Evropske unije, ki je nastal iz pomnilnikov prevodov, ustvarjenih med prevajanjem na Službi Vlade RS za evropske zadeve. Angleško-slovenski korpus vsebuje v času pisanja okrog 970.000 prevodnih enot (več kot 34 milijonov besed), nemško-slovenski korpus vsebuje več kakor 30.000 prevodnih enot (približno 1 milijon besed), francosko-slovenski korpus pa vsebuje več kakor 5.000 prevodnih enot (več kot 200.000 besed).
- **TRANS.** Ta korpus je nastal v okviru več študentskih projektov na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani. Iskanje po njem je mogoče preko skupnega iskalnega vmesnika s korpusom IJS-ELAN.

Za potrebe terminologije se uporabljajo specializirani korpusi, se pravi zbirke besedil z določenega strokovnega področja. Za nekatere metode pri ugotavljanju terminološkosti potrebujemo za primerjavo še splošnojezikovni, v idealnem primeru referenčni korpus jezika. Dvojezično terminografsko delo je močno olajšano, če imamo na razpolago vzporedni korpus, čeprav lahko za številne potrebe zadošča tudi primerljivi korpus. Strokovni jezik je v veliki meri vezan na pisni medij, zato govornih vzorcev navadno ne vključujemo. Terminologija je tudi pretežno usmerjena v raziskovanje in opisovanje sodobnega jezika, zato so specializirani korpusi za terminološke namene sinhroni in težijo k vključevanju čim sodobnejših besedil.

V zadnjem času je pri nas nastalo nekaj specializiranih korpusov, ustvarjenih za namene terminografije:

- **DSI.** To je korpus s področja informatike, ki vsebuje zbornike srečanja Dnevi slovenske informatike od leta 2003 do vključno 2007, vsako leto pa se dopolni s svežim zbornikom. Trenutno vsebuje 1,2 milijona besed in je prosto dostopen za iskanje.¹⁷ Ker je informatika s terminološkega stališča izredno živahna veda, je korpus dragocena podpora pri terminografskem projektu Islovarja¹⁸ (Erjavec in Vintar 2004).
- **Korpus slovenskih vojaških besedil** (Gorjanc in Logar 2007). Ta korpus vsebuje 5,5 milijonov besed iz različnih serijskih in drugih publikacij s področja vojaštva (Naša obramba, Slovenska vojska itd.), uporablja pa se pri gradnji

¹⁶ <http://nl2.ijs.si/corpus/index-bi.html>

¹⁷ <http://nl2.ijs.si/index-mono.html>

¹⁸ <http://www.islovar.org>

novega štirijezičnega vojaškega slovarja, ki nastaja pod okriljem sodelavcev ljubljanske Fakultete za družbene vede s partnerji.

- **KoRP.** KoRP je sinhroni enojezični korpus pisnih besedil odnosov z javnostmi. Obsega 1.824.699 besed in zajema besedila iz obdobja od leta 1994 do leta 2007. Je rezultat 1. faze projekta Slovar slovenskega izrazja odnosov z javnostmi, katerega izvajalec je Center za družboslovnoterminološko in publicistično raziskovanje Fakultete za družbene vede, naročnik in glavni financer pa podjetje Pristop, d. o. o. Projekt se je začel leta 2006. Pobuda za izdelavo terminološkega slovarja odnosov z javnostmi je prišla s strani dr. Dejana Verčiča, vodenje projekta je prevzela dr. Monika Kalin Golob, korpus pa je kot del svoje doktorske disertacije izdelala mag. Nataša Logar. Zbiranje besedil je potekalo v dveh fazah: (a) od aprila do maja 2006 in (b) od decembra 2006 do marca 2007. Računalniško obdelavo in jezikoslovno označitev je maja 2007 izvedlo podjetje Amebis, d. o. o., ki je omogočilo tudi spletni dostop.¹⁹

7.1.2 Reprezentativnost

Jedro vseh korpusnih pristopov je sklepanje o jeziku na podlagi vzorca, to se pravi, da mora korpus predstavljati reprezentativen vzorec jezika ali strokovnega jezika, ki ga opazujemo. O tem, kdaj lahko korpus upravičeno obravnavamo kot reprezentativen vzorec opazovane populacije, v našem primeru jezika, je razpravljalo že mnogo avtorjev. Predvsem pri raziskavah jezikovnih zvrsti in jezikovne norme je uravnotežena sestava korpusa ključni dejavnik, saj so v nasprotnem primeru pogostostna razmerja v korpusu zamaknjena in onemogočajo vrednotenje statističnih rezultatov.

Dva pomembna vidika reprezentativnosti pri gradnji korpusov povzame Biber (1993):

- **Raznosterost**

Analize kažejo, da se raba leksikalnih, slovničnih in diskurzivnih prvin močno spreminja glede na jezikovno zvrst, zato je raznosterost s smislu zastopanosti čim večjega števila različnih zvrsti in registrov ključnega pomena. Pravzaprav se kaže, da je pojem splošni jezik povsem abstrakten, saj ima vsaka besedilna vrsta oziroma vsak register lastne vzorce jezikovne rabe. Raznosterost pomeni tudi ustrezno vključevanje dialektalnih ali regionalnih jezikovnih zvrsti, pa tudi uravnoteženost v smislu tematskih področij.

¹⁹ <http://www.korp.fdv.uni-lj.si/>

- **Velikost**

Čeprav se večina razprav o sestavi korpusov osredotoča na skupno velikost korpusa v besedah, vprašanje o velikosti vključuje tudi število besedil iz posamezne zvrsti, število vzorcev iz posameznega besedila in število besed v posameznem vzorcu.²⁰ Biber v eni svojih vplivnih študij pokaže, da je za ugotavljanje lastnosti določene jezikovne zvrsti 10 besedil že dovolj velik vzorec, od vsakega besedila pa zadošča naključno izbran vzorec v dolžini 1000 besed. Po drugi strani je jasno, da je za leksikografske namene potrebno ogromno besedilnega materiala; tu se številke danes gibljejo od 100 milijonov besed naprej, zgornje meje pa pravzaprav ni.

O načelih gradnje specializiranih korpusov je bilo povedanega mnogo manj, predvsem pa so kriteriji tu močnejše vezani na namembnost korpusa. Gorjanc to situacijo povzema takole (Gorjanc 2002: 79):

Korpusi strokovnih jezikov so se pojavili precej kasneje kot splošni referenčni, zato se načela za njihovo gradnjo zares šele oblikujejo; glede splošnih postopkov gradnje tudi pri njih veljajo ista izhodišča, a se glede na to, da zajemajo le jezik v točno določeni funkciji, dopolnjujejo in na novo premišljajo. Zaradi dinamike znanstvenega in tehnološkega razvoja pa pomenijo tisto osnovo, ki bo omogočala sprotno spremljanje jezikovnega dogajanja na strokovnih področjih in delovala predvsem na ravni terminološkega usklajevanja, ki je zaradi dinamike razvoja vse težje obvladljivo; pomembni so tudi kot izjemno dragocen vir podatkov o trenutnem vedenju.

Poudariti je treba, da je gradnja specializiranih korpusov za namene terminografije v luči zgornjih vidikov podrejena nekoliko drugačnim načelom. Korpus strokovnih besedil naj bi čim bolj predstavljal določeno stroko in vseboval kar največ strokovnih izrazov.

Prva težava se pojavi že ob opredelitvi strokovnega področja. Danes strok ne moremo več jasno deliti na naravoslovno-tehnične in družboslovno-humanistične, niti ni mogoče posameznih področij povsem razmejiti od ostalih in jih prirediti zgolj eni veji znanosti. Prekrivanje med področji in strokami ter prehajanje strokovnih izrazov v splošni jezik se med drugim kaže v vse večji večpomenskosti posameznih izrazov, če jih obravnavamo izven njihovega konteksta.

Še mnogo bolj pa se to prekrivanje pokaže v dejanskih besedilih, ki naj bi služila kot vzorec za določeno področje. Skoraj vsa besedila namreč tu in tam posegajo na druga

²⁰ Predvsem korpusi prve generacije, na primer Brown ali LOB, niso vključevali besedil v celoti, ampak je bil iz vsakega besedila izbran vzorec določene velikosti. Tudi danes se za zagotavljanje reprezentativnosti ponekod odločimo za krajšanje besedil ali za izbor naključno izbranega vzorca določene dolžine.

področja, govorijo o drugih temah in uporabljajo »gostujoče« izrazje. Pri korpusni obdelavi terminološkega inventarja se navadno osredotočamo na določeno področje, zato so izrazi z drugih področij pri rezultatih nezaželeni šum, čeprav so morda informativni za besedilo. Korpusna terminografija tako zahteva besedila, ki so čim tesneje povezana s področjem in ne vsebujejo preveč ekskurzov na druga področja.

Podobni argumenti veljajo tudi za vprašanje časovnega razpona besedil. Ker so strokovna področja danes podvržena izredno hitremu razvoju, se tudi besedišče strok s časom pospešeno spreminja. Velik časovni razpon besedil zato pomeni slabše rezultate pri avtomatski obdelavi terminologije, saj moramo računati z nedoslednimi izrazi, razlikami pri zapisu terminov in podobno. Po drugi strani pa so prav terminološke variacije eden od vidikov, ki morajo biti v terminografski zbirki ustrezno zaobjeti, če želimo doseči čim širšo uporabnost nastalega jezikovnega vira.

Zelo pomembna je tudi velikost, saj za terminološke metode veljajo podobna načela kot za korpusno leksikografijo. Za razpoznavanje tipičnih terminoloških vzorcev je potrebna večkratna pojavitev izraza, kar zagotavlja le ustrezna velikost, obenem pa bomo z večjim vzorcem besedil sposobni pokriti več izrazja izbrane stroke.

Kadar je naš namen iskanje izrazja v dvojezičnem okolju na podlagi vzporednega korpusa, se ob zgornjih vprašanjih pojavijo še specifični pomisleki v zvezi s prevodom besedil. Pri snovanju vzporednega korpusa je najprej pomembna odločitev, ali je smer prevoda pomembna ali ne. Če nameravamo korpus uporabljati za prevodoslovnih raziskave, na primer o prevajalskih strategijah ali značilnostih prevedenih besedil, mora biti korpus glede smeri prevoda opredeljen. To pomeni, da izberemo le en jezik izvirnika in en jezik prevoda, vsa vzporedna besedila v korpusu pa so torej prevodi v isti jezik. Kadar korpus vsebuje besedila obeh prevodnih smeri, kot je to na primer pri korpusih IJS-ELAN in TRANS, moramo smer prevoda pri vsakem besedilu označiti, tako da je iskanje po korpusu mogoče omejiti na določeni prevodni par.

Pri terminološkem delu to vprašanje prav tako ni nepomembno, predvsem kar se tiče uporabe dobljenih rezultatov. Terminolog ali prevajalec, ki bi želel uporabiti vzporedni korpus kot vir prevodnih ustreznih, mora vsekakor poznati sestavo korpusa in naravo besedil v njem, saj bo drugače težko vrednotil status najdenih izrazov in njihovih možnih ustreznih.

Zadnje pomembno vprašanje v zvezi z reprezentativnostjo pa je kakovost besedil in njihovih prevodov, saj je od tega še najbolj odvisna tudi kakovost pridobljenega terminološkega gradiva. Prevodi besedil z določenega področja, kjer si prevajalci niso prizadevali za terminološko doslednost, kjer so besedilo terminološko osiromašili

oziroma izbirali razlagalne rešitve, za terminografske namene niso primerni. Po drugi strani pa nam tudi pri takih besedilih avtomatske metode pomagajo ugotavljati nedoslednosti in terminološke variacije.

7.1.3 Homogenost

Na tem mestu se nam zdi smiselno vključiti kratko razpravo o homogenosti specializiranih korpusov. V prejšnjem razdelku smo pokazali, da za namene terminologije raznoterost ni nujno zaželena lastnost, nasprotno pa bi si želeli, da je korpus čimbolj homogen v smislu besedišča, ki ga predstavlja.

Kilgarrieff (2001) se prvi sistematično posveti vprašanju, kako primerjati korpusa med seboj in kako ugotavljati razlike med njimi, obenem pa raziskuje tudi vprašanje homogenosti korpusa. Glede na razmeroma dolgo zgodovino korpusnega jezikoslovja je kar neverjetno, da se tej temi prej nihče ni podrobneje približal.

Za raziskovanje podobnosti oziroma razlik med dvema korpusoma Kilgarrieff predlaga metodo primerjave pogostosti besed, ki temelji na Mann-Whitneyevem testu uvrstitev na lestvici pogostosti. Oba korpusa najprej razdelimo na vzorce enakih velikosti, tako da so pogostosti besed neposredno primerljive med vsemi vzorci. Test temelji na ničti hipotezi, in sicer da sta korpusa enaka oziroma da so vsi vzorci vzeti iz iste populacije. Če lahko dokažemo, da so pogostosti iz vzorcev prvega korpusa v povprečju višje ali nižje od pogostosti iz vzorcev drugega korpusa, lahko ničto hipotezo statistično zavržemo z znano mero zaupanja.

Kilgarrieff metodo uporabi za primerjavo korpusa Brown (ameriška angleščina) in LOB (britanska angleščina) in z njo identificira besede, ki so tipične za vsakega od korpusov. Obenem metodo primerja z drugimi znanimi metodami za iskanje ključnih besed, kot so t-test, MI ali X^2 , in ugotovi, da daje najboljše rezultate za ta namen.

Na podoben način lahko ugotavljamo tudi homogenost korpusa. Tu Kilgarrieff vprašanje homogenosti prevede na vprašanje podobnosti korpusa samemu sebi, za testiranje različnih cenilk pa vzpostavi kontrolno populacijo, za katero je podobnost znana. Svojo metodo zato poimenuje metoda korpusov znanih podobnosti oziroma *Known-Similarity Corpora*.

Kako simuliramo znano podobnost? Postopek je precej preprost. Vzamemo dva korpusa precej različnih jezikovnih zvrsti, A in B. Nato zgradimo skupino kontrolnih korpusov KK tako, da KK1 vsebuje 100% A, KK2 90% A in 10% B, KK3 80% A

in 20% B, KK4 70% A in 30% B in tako naprej. Zdaj lahko trdimo, da je KK2 bolj podoben korpusu A kot KK3, KK4 je bolj podoben korpusu B kot KK1 in podobno. S temi kontrolnimi skupinami lahko testiramo različne metode merjenja podobnosti med korpusi.

Da bi merili homogenost enega korpusa, potrebujemo naslednje korake:

- korpus razdelimo na enako velike »porcije«,
- ustvarimo dva podkorpusa, tako da porcije naključno priredimo prvemu ali drugemu podkorpusu,
- izmerimo podobnost med korpusoma,
- postopek ponavljamo z različnimi razdelitvami porcij,
- izračunamo povprečje in standardno deviacijo vseh ponovitev.

Primerjava različnih statistik pri tej metodi pokaže, da je najzanesljivejša X^2 , ki se obnese bolje kot Spearmanov korelacijski koeficient uvrstitev in precej bolje kot različne metode z uporabo entropije.

X^2 je test za preverjanje neodvisnosti spremenljivk, kjer izhajamo iz ničte hipoteze. Na splošno se v korpusnem jezikoslovju pokaže, da ta statistika ni uporabna za preverjanje jezikoslovnih hipotez, kajti korpus ni nikdar naključni vzorec besed. Vendar pa v tem primeru Kilgarrieff predlaga nekoliko drugačno rabo testa, ki se obnese.

Rezultati pokažejo, da ta statistika dobro predstavi razlike v pogostosti besed med dvema korpusoma, vrednost cenilke pa postopoma narašča z naraščajočo pogostostjo besede. To ustreza intuitivni predpostavki, da so bolj pogoste besede boljše merilo različnosti oziroma podobnosti korpusov kot manj pogoste. Za preskus homogenosti strokovnega korpusa bi opisano metodo uporabili tako, da bi test enkrat izvedli s korpusoma različnih področij, drugič pa z istim korpusom, naključno razdeljenim na dva dela.

7.1.4 WebBootCaT

Tudi gradnjo specializiranega korpusa je mogoče avtomatizirati. Taka storitev je WebBootCaT,²¹ stroj za hitro gradnjo specializiranih korpusov za prevajalske name-
ne (Baroni in dr. 2006). Po predhodni prijavi, ki je brezplačna za 30 dni, se uporabniku odpre spletni formular, kamor vnese ključne besede. Te se nato uporabijo kot »šeme« pri brskanju po spletu, uporabnik pa določi tudi jezik besedil, ki naj se zberejo v korpusu. WebBootCaT nato uporabi spletni iskalnik (Google ali Yahoo!)

²¹ <http://sketchengine.co.uk/auth/wbc/>

za zbiranje spletnih strani, pri tem pa sproti odstranjuje podvojene strani ali strani, ki ne vsebujejo dovolj besedila.

Sledi faza samodejnega čiščenja, ki je izredno dragocena, saj iz spletnih dokumentov pobriše ostanke kode html, java in skriptov, pa tudi tisto besedilo, ki je vezano na spletno navigacijo. Besedila se tudi tokenizirajo, se pravi razčlenijo na stavke ter naprej na besedne oblike in ločila.

Slika 11: Spletni servis za gradnjo korpusov BootCaT



BootCaT
www version

user: Špela Vintar, free space: 436774 tokens

This application uses the [Yahoo! Web Services](#).

Please make sure you have JavaScript enabled in your web browser.

Seed words ribištvo ribarjenje "ribolovne cone" staleži
Use space as separator. Enclose multiwords expressions into quotes ("").

Upload seed words in a plain text file -- one expression per line.

Language Slovenian
Select the language of the corpus to be built.

Select URLs ☐
Check this box if you would like to manually filter found URLs before the webpages will be downloaded.

Tag corpus ☐
Your corpus will be POS-tagged and lemmatized using the [TreeTagger](#). Following languages are currently supported: Bulgarian, Dutch, English, French, German, Italian, Russian, Spanish. This option has no effect if used with any other languages.

Če gradimo korpus za enega od bolj podprtih jezikov (angleščina, bolgariščina, francoščina, italijanščina, nemščina, nizozemščina, ruščina, španščina), je na voljo tudi oblikoslovno označevanje s označevalnikom TreeTagger. Za nekatere od zgornjih je-

zikov so na voljo tudi referenčni korpusi, ki nam omogočajo samodejno izdelavo seznamov ključnih besed za specializirani korpus.

Ko je korpus na podlagi vnešenih ključnih besed zgrajen (za korpus v slovenščini velikosti 70.000 besed program porabi približno dve minuti), ga lahko uporabljamo prek iskalnika SketchEngine (Kilgariff in dr. 2004), lahko pa si ga tudi naložimo na lastni računalnik v goli ali tokenizirani obliki. Sicer je uporabniški prostor na strežniku WBC-ja omejen na 500.000 besed, proti plačilu pa si lahko to omejitev tudi povečamo.

Iskanje prek SketchEngina nam seveda ponuja številne prednosti, saj lahko uporabimo vgrajene statistike za iskanje kolokacij, v okviru zgrajenega korpusa določamo podkorpuse, gradimo besedne seznime in podobno. Pomanjkljivost WebBootCaT-a je morda le, da je pri določanju jezika besedila prav tako nenatančen kot spletni iskalniki, zato se nam v korpus mimogrede prikradejo tudi besedila v angleščini. Za obsežnejše terminološke raziskave pa nam navadno ne zadoščajo besedila s spleta, zato moramo uporabiti druge načine zbiranja in obdelave.

7.2 OBDELAVA KORPUSA

Ko smo ustvarili specializirani korpus, potrebujemo programsko infrastrukturo za njegovo obdelavo, iskanje in preverjanje hipotez. Koraki obdelave besedil so odvisni od razpoložljivih sredstev in namena uporabe korpusa, skoraj vedno pa vključujejo vsaj čiščenje besedil in pretvorbo v enotni zapis. Če gradimo vzporedni korpus, je pomemben korak stavčna poravnava besedil, za naprednejše iskanje in morebitno samodejno luščenje izrazja pa je neizogibno tudi oblikoskladenjsko označevanje.

7.2.1 Označevanje

Raven označenosti je izredno pomembna in pogojuje izbor metod za iskanje po korpusu, pa tudi za morebitno samodejno luščenje izrazja. Na spletu so brezplačno na voljo označevalniki za številne jezike, enega boljših seznamov prosto dostopnih označevalnikov vzdržuje stanfordska univerza na spletišču Statistical natural language processing.²²

Za slovenščino in druge pregibne jezike je poleg oblikoslovnega označevanja pomembna tudi lematizacija. Sodelavci Odseka za tehnologije znanja Instituta Jožef

²² <http://www-nlp.stanford.edu/links/statnlp.html#Taggers>

Stefan so izdelali spletni lematizator, s katerim je možno označiti krajša besedila.²³ Storitev ponuja dva algoritma, RDR in CLOG, pri čemer slednji vključuje jezikovne modele za slovenščino, angleščino, romunščino, češčino, estonščino in madžarščino, poleg lem pa izpiše tudi oblikoskladenjske oznake.

Tabela 5 podaja primer rezultatov obeh algoritmov za spodnji besedilni odsek, napa-ke so označene s krepkim tiskom:

Monitoring ribolovnih virov s pridneno vlečno mrežo je biološko raziskovanje, ki v sloven-skem morju poteka od leta 1995. Vzorčenje poteka z najetim ribiškimi plovilom, ki štirikrat letno opravi standardizirane ribolove s pridneno vlečno mrežo na treh sistematično določenih mestih v slovenskem morju.

Tabela 5: Spletna lematizacija za slovenščino

Besedna oblika	RDR	CLOG
Monitoring	Monitoring	monitoring
ribolovnih	riboloven	riboloven
virov	vir	vir
s	s	s
pridneno	pridnen	pridnen
vlečno	vlečen	vlečen
mrežo	mreža	mreža
je	biti	biti
biološko	biološki	biološki
raziskovanje	raziskovanje	raziskovanje
ki	ki	ki
v	v	v
slovenskem	slovenski	slovenski
morju	mor	morje
poteka	potekati	potek
od	od	od
leta	leto	leto
Vzorčenje	Vzorčenje	vzorčenje

²³ <http://nl2.ijs.si/analyze/>

poteka	potekati	potekati
z	z	z
najetim	najet	najet
ribiškim	ribiški	ribiški
plovilom	plovilo	plovilo
ki	ki	ki
štirikrat	štirikrat	štirikrat
letno	leten	leten
opravi	oprava	opraviti
standardizirane	standardizirati	standardiziran
ribolove	ribolov	ribolov
s	s	s
pridneno	pridnen	pridnen
vlečno	vlečen	vlečen
mrežo	mreža	mreža
na	na	na
treh	trije	trije
sistematično	sistematičen	sistematičen
določenih	določen	določen
mestih	mesto	mesto
v	v	v
slovenskem	slovenski	slovenski
morju	mor	morje

7.2.2 Obdelava korpusa z orodjem Wordsmith Tools

Pri korpusnem terminološkem delu se srečujemo z veliko količino besedilnih podatkov, iz katerih bi želeli izluščiti čimveč znanja o raziskovanem specializiranem področju. Vprašanja, ki si jih utegnemo zastaviti, so naslednja:

- Kako dobro moj korpus predstavlja izbrano področje? Ali je njegova sestava preveč heterogena? Ali vsebuje dovolj besedil?
- Kateri izrazi se v korpusu najpogosteje pojavljajo? Kateri najmanj pogosto?

- Katere specializirane izraze vsebuje korpus?
- Kaj pomeni določeni specializirani izraz?
- Kako se določeni izraz prevaja?
- Kako se določeni izraz uporablja v kontekstu?
- Katere izraze je smiselno vključiti v terminološko bazo?

Na vsa ta vprašanja seveda ni enoličnega odgovora, a s pomočjo ustreznih orodij je tudi brez programerskega znanja in posebnih predpriprav mogoče obdelovati in raziskovati velike količine besedil. Če ob tem dobro poznamo tudi korpusne metode, ki jih imamo na voljo, se lahko zgornjim ciljem približamo z izbiro ustreznega orodja za raziskovanje korpusnega gradiva. Eno takih orodij je Wordsmith Tools,²⁴ ki ga v nadaljevanju uporabimo za vse podane primere, na razpolago pa je še precej sorodnih proizvodov, nekaterih tudi brezplačnih, ki povečini omogočajo statistične obdelave.

Wordsmith Tools ponuja skupek programov za izdelavo besednih seznamov, seznamov ključnih besed in za konkordančno iskanje, ob tem pa aplikacija vključuje še orodja za kosanje in združevanje datotek, iskanje in zamenjavo nizov v korpusu, iskanje besedil na spletu itd. Program je primeren za delo z enojezičnimi korpusi, saj nima funkcij za vzporedne konkordance ali dvojezično iskanje. Vhodni podatki za delo z Wordsmithom morajo biti pretvorjeni v golo besedilo (.txt), program pa podpira različna znakovna kodiranja (ANSI, Windows, Unicode). Pri delu z označenimi besedilnimi formati, denimo korpusi v XML, je Wordsmith okoren in omogoča le zelo primarne načine obdelave. Program ni brezplačen, a cena zanj tudi ni vrtoglava, demo različico pa si je mogoče naložiti z njegove domače strani in ponuja preskus vseh funkcij, a z omejenim številom izpisanih vrstic.

Ob zagonu programa se najprej odpre Wordsmithov kontrolnik, ki ponuja glavne tri programske komponente, **Concord**, **Keywords** in **Wordlist**. V nadaljevanju na kratko predstavimo vse tri z vidika uporabnosti za terminološke raziskave.

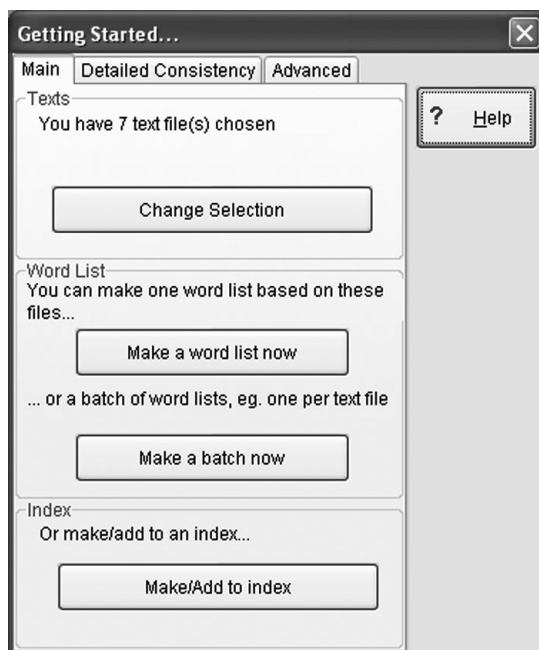
7.2.2.1 Besedni seznam

Orodje Wordlist nam omogoča vertikalni vpogled v korpus, se pravi vpogled v besedni inventar izbranih besedil. Za izdelavo osnovnega besednega seznama s klikom na zeleni gumb priključimo okno **Getting started**, kjer v prvi fazi izberemo besedila (Slika 12). Program omogoča delo s številnimi datotekami in velikimi besedilnimi zbirkami, če so ta le vsa v enakem kodiranju in shranjena kot golo besedilo. Ko smo izbrali eno ali več besedil, imamo na voljo tri možnosti, in sicer izdelavo enega be-

²⁴ <http://www.lexically.net>

sednega seznama za celoten korpus (**Make a wordlist now**), izdelavo več seznamov (**Make a batch now**) in izdelavo besednega indeksa (**Make/Add to index**). Slednja možnost je pomembna predvsem za analizo besednih skupkov (clusters).

Slika 12: Začetek dela z Wordlistom



Z izbiro prve možnosti program prečeše izbrane datoteke in prikaže besedni seznam, urejen po pogostosti. V prvi fazi ukvarjanja s terminologijo v specializiranem korpusu si bomo verjetno ogledali prav tega. Že iz pregleda prvih nekaj sto besed na pogostostnem seznamu si ustvarimo vtis o tem, kako dobro naš korpus predstavlja izbrano področje. Da je besedni seznam bolj pregleden, je smiselno pri njegovi izdelavi uporabiti t. i. seznam praznih besed. To je seznam tistih besednih enot, ki nas v prvi fazi terminoloških raziskav ne zanimajo, se pravi veznikov, členov, predlogov, zaimkov in pomožnih glagolov — te bi sicer zaradi svoje pogostosti zavzele zgornji del pogostostnega seznama.

Če želimo pri izdelavi besednega seznama uporabiti seznam praznih besed, to možnost aktiviramo v nastavitvah Wordlista (Settings – Stop-, Lemma- & Matchlists). Besedni seznam moramo po vsaki spremembi nastavitve izdelati znova. Primer pogostostnega seznama iz zbirke besedil o davčni zakonodaji s področja transfernih cen,

iz katerega so izločene prazne besede, kaže Tabela 6. Višaj na prvem mestu pomeni številčne pojavnice, ki so združene pod eno različnico.

Tabela 6: Pogostostni seznam z izločenimi praznimi besedami

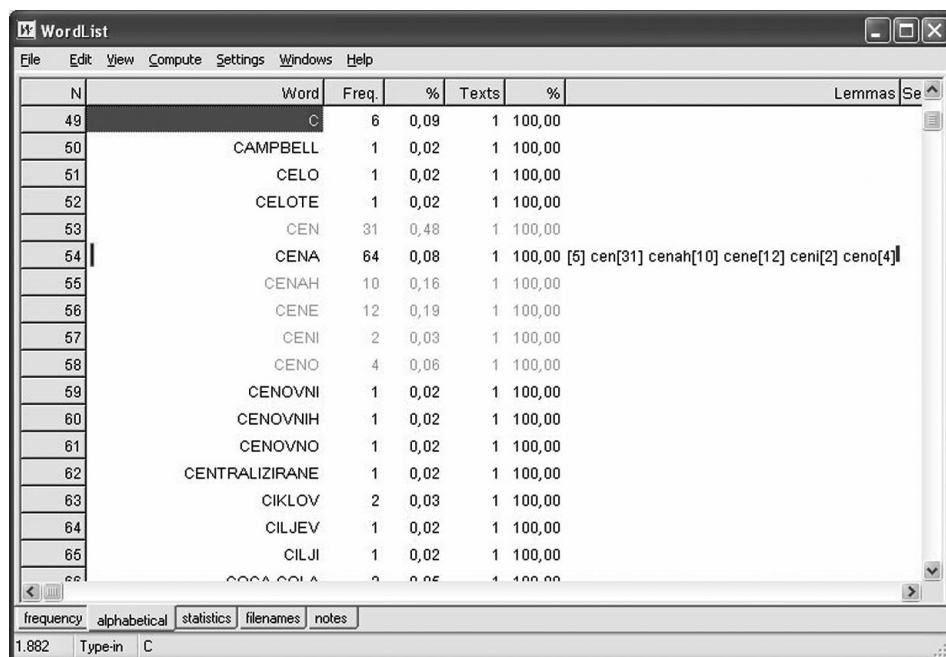
	Različnica	Pogostost	Rel. pogostost
1	#	438	6,7991
2	PODJETJA	36	0,5588
3	ZNAMKE	35	0,5433
4	BLAGOVNE	34	0,5278
5	PRIMERLJIVOSTI	33	0,5123
6	TRANSFERNIH	32	0,4967
7	CEN	31	0,4812
8	SREDSTVA	28	0,4346
9	POTREBNO	26	0,4036
10	PODJETJE	22	0,3415
11	SREDSTEV	22	0,3415
12	POSILI	18	0,2794
13	PRIMERLJIVOST	18	0,2794
14	NEOPREDMETENA	17	0,2639
15	ZNAMKA	17	0,2639
16	ANALIZA	16	0,2484
17	BLAGOVNA	16	0,2484
18	PODATKOV	16	0,2484
19	TRANSAKCIJE	16	0,2484
20	VEČ	16	0,2484
21	METODE	15	0,2328
22	STORITVE	14	0,2173
23	TVEGANJ	14	0,2173
24	LASTNINE	13	0,2018
25	NEOPREDMETENIH	13	0,2018

Pogostostni seznam je vstopna točka za terminološko delo s korpusom. Ročni pregled vsaj prvih nekaj sto različnic na seznamu nam poda prvi vtis o tem, kako dobro naš korpus pokriva izbrano področje in ali je zastopano besedišče pretežno s tega področja.

Pogostost je eden od kriterijev terminološkosti, čeprav seveda ne edini ali najboljši. Iz zgornje tabele denimo lahko utemeljeno sumimo, da so *podjetje*, *primerljivost*, *sredstva*, *transakcija*, *tveganje*, *lastnina* itd. na področju transference cen termini ali deli terminov, s seznama pa lahko razberemo tudi nekaj večbesednih terminoloških kandidatov: *blagovna znamka*, *transferne cene*, *neopredmetena sredstva*. Če želimo že v fazi pregledovanja enobesednega seznama kako različnico preveriti v sobesedilu, je iz Wordlista mogoč neposredni prekop v program Concord, kjer se za izbrano različnico izdela konkordanca.

Wordlist omogoča različne načine urejanja seznama, po pogostosti, abecedno, po dolžini besed, po končnicah z odzadnjim urejanjem. Iz abecedno urejenega besednega seznama lahko razberemo, kateri leksemi se v korpusu pojavljajo v številnih izpeljanih oblikah in kateri ne; besedotvorno izrazito razvejani leksemi so pogosto indikator terminološkega gnezda. V abecedno urejenem seznamu se je mogoče lotiti tudi ročne lematizacije, se pravi pripisovanja posameznih besednih oblik določeni osnovni obliki ali lemi. Slika 13 kaže primer ročne lematizacije besede *cena* in njenih oblik. Besedne oblike z miško povlečemo na lemo, program pa samodejno sešteje pogostosti in nato besedni seznam tudi ponovno razvrsti.

Slika 13: Ročna lematizacija v Wordlistu



N	Word	Freq.	%	Texts	%	Lemmas
49	C	6	0,09	1	100,00	
50	CAMPBELL	1	0,02	1	100,00	
51	CELO	1	0,02	1	100,00	
52	CELOTE	1	0,02	1	100,00	
53	CEN	31	0,48	1	100,00	
54	CENA	64	0,08	1	100,00	[5] cen[31] cenah[10] cene[12] ceni[2] ceno[4]
55	CENAH	10	0,16	1	100,00	
56	CENE	12	0,19	1	100,00	
57	CENI	2	0,03	1	100,00	
58	CENO	4	0,06	1	100,00	
59	CENOVNI	1	0,02	1	100,00	
60	CENOVNIH	1	0,02	1	100,00	
61	CENOVNO	1	0,02	1	100,00	
62	CENTRALIZIRANE	1	0,02	1	100,00	
63	CIKLOV	2	0,03	1	100,00	
64	CILJEV	1	0,02	1	100,00	
65	CILJI	1	0,02	1	100,00	

frequency alphabetical statistics filenames notes

1.882 Type-in C

Z Wordlistom lahko izdelujemo tudi sezname večbesednih enot, kar je za terminološke raziskave zelo uporabno. V ta namen moramo prej izdelati indeks, nato pa v nastavitvah določimo dolžino večbesedne enote (med 2 in 12 besed) ter najmanjšo zahtevano pogostost. Slika 14 kaže seznam dvobesednih enot, med katerimi je znova kar nekaj terminološko zanimivih.

Slika 14: Dvobesedne enote v orodju Wordlist

N	Word	Freq.	%	Texts	%	emr
1	BLAGOVNE ZNAMKE	33	0,51	1	100,00	
2	TRANSFERNIH CEN	22	0,34	1	100,00	
3	NEOPREDMETENA SREDSTVA	17	0,26	1	100,00	
4	BLAGOVNA ZNAMKA	16	0,25	1	100,00	
5	BLAGOVNIH ZNAMK	11	0,17	1	100,00	
6	FUNKCIJSKA ANALIZA	10	0,16	1	100,00	
7	OBLIKOVANJA TRANSFERNIH	10	0,16	1	100,00	
8	TRANSFERNIH CENAH	10	0,16	1	100,00	
9	LASTNINE DOBRIN	9	0,14	1	100,00	
10	NEOPREDMETENIH SREDSTEV	9	0,14	1	100,00	
11	NEODVISNEGA TRŽNEGA	7	0,11	1	100,00	
12	OBLIKOVANJE TRANSFERNIH	7	0,11	1	100,00	
13	FUNKCIJ SREDSTEV	6	0,09	1	100,00	
14	FUNKCIJSKO ANALIZO	6	0,09	1	100,00	
15	NEVTRALNIH IZIDOV	6	0,09	1	100,00	
16	POSLOVNE STRATEGIJE	6	0,09	1	100,00	
17	POVEZANA PODJETJA	6	0,09	1	100,00	

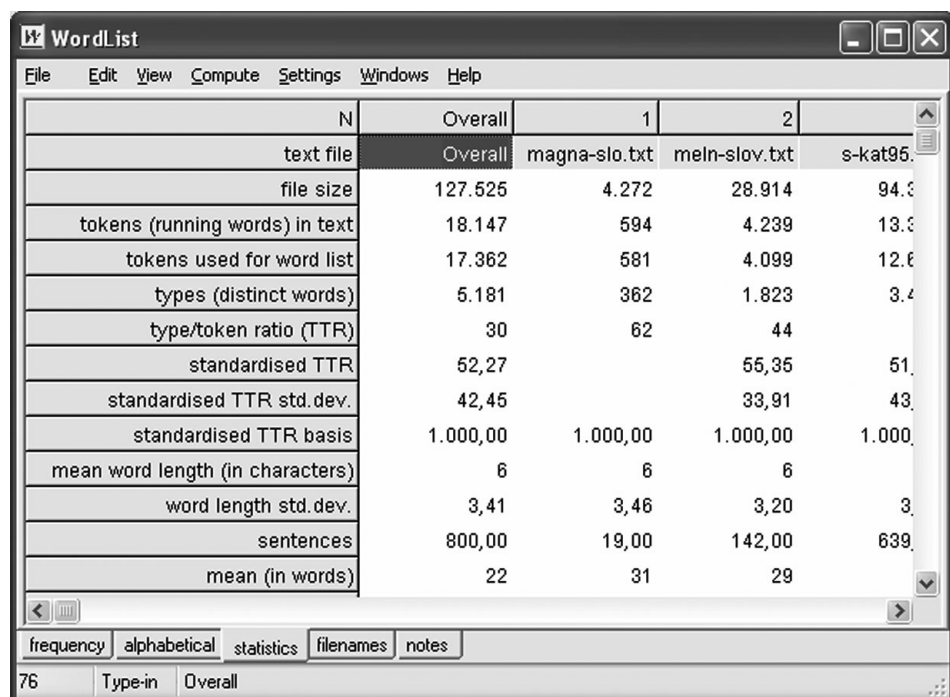
frequency alphabetical statistics filenames notes

123 Type-in STOPNJE PRIMERLJIVOSTI

Orodje Wordlist pa pod jezičkom Statistics ob vsaki izdelavi besednega seznama predstavi tudi osnovne podatke o korpusu. Ta osebna izkaznica korpusa za celotno zbirko in za vsako posamezno besedilo podaja velikost korpusa v bajtih, pojavnih in različnih, stavkih in odstavkih. Razmerje med pojavnimi in različnimi je izraženo z absolutnim in standardiziranim TTR-jem (angl. *type/token ratio*), kar nam omogoča vpogled v bogatost besedišča korpusa in posameznih besedil. Nadalje statistika poda povprečno dolžino odstavka, povedi in besede, preštete pa so tudi besede po dolžini v znakih (Slika 15).

Wordlist nam pri terminološkem delu pomaga predvsem v fazi izdelave geslovnika, saj je besedne sezname mogoče poljubno urejati, razvrščati, shranjevati in izvažati v .txt ali Excelovo tabelo. Seveda pa resno terminografsko delo pomeni pregledovanje prav vseh različnic na seznamu, saj — kot rečeno — pogostost ni nujni pogoj za terminološkost; nasprotno se številni visoko specializirani termini v besedilih pojavljajo redko ali celo samo enkrat. Prav tako je za korpusno terminografijo nujno zavedanje, da še tako obsežen in reprezentativen korpus ne more zajeti prav vsega izrazja določenega področja. Še posebej za ustrezno predstavitev terminoloških variacij ali dvojnic, pa tudi za dopolnjevanje iz korpusa zajetih pojmovnih sistemov, je navadno potrebno ciljno nadgrajevanje korpusnega gradiva z drugimi viri.

Slika 15: Wordsmith — osnovna statistika korpusa



The screenshot shows the WordList application window. The menu bar includes File, Edit, View, Compute, Settings, Windows, and Help. The main window displays a table of statistics for a text file. The table has five columns: N, Overall, 1, 2, and s-kat95. The rows list various statistical measures such as file size, tokens, types, TTR, and mean word length. At the bottom, there are tabs for frequency, alphabetical, statistics, filenames, and notes. The status bar at the very bottom shows '76 Type-in Overall'.

	N	Overall	1	2	s-kat95
text file		Overall	magna-slo.txt	meIn-slov.txt	s-kat95
file size		127.525	4.272	28.914	94.3
tokens (running words) in text		18.147	594	4.239	13.3
tokens used for word list		17.362	581	4.099	12.6
types (distinct words)		5.181	362	1.823	3.4
type/token ratio (TTR)		30	62	44	
standardised TTR		52,27		55,35	51
standardised TTR std.dev.		42,45		33,91	43
standardised TTR basis		1.000,00	1.000,00	1.000,00	1.000
mean word length (in characters)		6	6	6	
word length std.dev.		3,41	3,46	3,20	3
sentences		800,00	19,00	142,00	639
mean (in words)		22	31	29	

frequency alphabetical statistics filenames notes

76 Type-in Overall

7.2.2.2 Konkordance

Raziskovanje frazeološkega vedenja posameznih terminološko zanimivih leksikalnih enot je najenostavneje prek konkordance. Za to ponuja Wordsmith orodje Concord, v katerem prav tako najprej izberemo besedila, nato pa definiramo iskalni pogoj.

Pri iskanju program omogoča uporabo nadomestnega znaka * in poševnice za zapis alternativnih nizov. Tako denimo iskalni pogoj *profit*/dobič** poišče vse pojavitve bodisi besed, ki se pričnejo s *profit*-, bodisi besed, ki se pričnejo z *dobič*-. Urejanje konkordance po levem ali desnem okolju nam že na prvi pogled razkrije kandidate za večbesedne termine in kolokacije, npr. *neto profitna marža*, *porazdelitev dobička*, *metoda porazdelitve dobička*, *čisti dobiček* itd.

Medtem ko za odkrivanje večbesednih terminov zadostuje opazovanje neposredne okolice jedrnega izraza, pa se kolokacije lahko pojavljajo tudi v precejšnji oddaljenosti od jedra. Wordsmith ponuja samodejno računanje kolokatorjev z različnimi statistikami (specifična vzajemnost — SMI, MI3, logaritem razmerij verjetja — LL, vrednost Z). Poleg tabelarnega izpisa kolokatorjev si je s programom mogoče ogledati še večbesedne skupke, ki se pojavljajo kjerkoli v izpisani konkordanci, ne le v kombinaciji z jedrno besedo. Funkcija Nariši (Plot) predstavi pojavitve iskanega niza na linearni osi, kjer vsaka pokončna črtica pomeni eno pojavitev. To nam omogoča vpogled v razporeditev iskanega niza po besedilni zbirki.

7.2.2.3 Ključne besede

Še korak naprej v smeri izbora pomembnih terminoloških enot pa naredimo, če primerjamo pogostostni besedni seznam specializiranega korpusa s pogostostnim seznamom referenčnega korpusa, saj je na ta način mogoče za vsako besedno obliko ali lemo iz specializiranega korpusa izračunati njeno ključnost. V programu Wordsmith Tools je za to na voljo komponenta Keywords. Če se beseda *a* v specializiranem korpusu *S* velikosti N_S pojavi s pogostostjo f_S , hkrati pa se beseda *a* v referenčnem korpusu *R* velikosti N_R pojavi s pogostostjo f_R , je razmerje njunih relativnih pogostosti hkrati merilo ključnosti besede *a*:

$$k(a) = \frac{\frac{f_S}{N_S}}{\frac{f_R}{N_R}}$$

Leksikalne enote, ki se pojavijo pri vrhu seznama ključnih besed in imajo hkrati visoko pogostost, navadno predstavljajo temeljne izraze določenega področja, zato lahko pri njih pričakujemo, da se bodo pojavljale v večbesednih terminih in tvorile terminološke kolokacije. V Tabeli 7 so navedene ključne besede iz majhnega korpusa farmacevtskih besedil, ki smo ga s programom Keywords primerjali z večjim korpusom splošnih besedil.

Tabela 7: Ključne besede iz korpusa farmacevtskih besedil

	Različnica	Pogostost	Rel. pogostost
1	#	632	4,7469
2	MG	112	0,8412
3	R	90	0,676
4	DELOVANJE	56	0,4206
5	TABLETE	49	0,368
6	UPORABA	48	0,3605
7	G	49	0,368
8	SESTAVA	41	0,3079
9	REGISTRACIJSKI	39	0,2929
10	STATUS	39	0,2929
11	DOZIRANJE	36	0,2704
12	INDIKACIJE	35	0,2629
13	KONTRAINDIKACIJE	35	0,2629
14	SREDSTVO	33	0,2479
15	ZDRAVLJENJE	32	0,2403
16	VSEBUJE	30	0,2253
17	UČINKI	30	0,2253
18	STRANSKI	29	0,2178
19	ZDRAVILO	28	0,2103
20	ML	27	0,2028
21	ZDRAVILA	26	0,1953
22	OL	26	0,1953
23	ZDRAVILNO	25	0,1878
24	LEKOVIT	24	0,1803
25	PRIPRAVKI	24	0,1803
26	POMOŽNO	24	0,1803
27	KAPSULE	23	0,1728
28	KAPLJIC	21	0,1577
29	OTROKOM	24	0,1803

30	MENTOKLAR	20	0,1502
31	DELUJE	20	0,1502
32	PRIPOROČAMO	20	0,1502
33	DIVERIN	20	0,1502
34	GEL	20	0,1502
35	KISLINE	20	0,1502
36	LASTNOSTI	21	0,1577
37	PRIMEREN	20	0,1502
38	TRIKRAT	22	0,1652
39	BOLEČINE	19	0,1427
40	ZA	171	1,2844

Za preiskovanje lastnih korpusov imamo torej na voljo programska orodja, ki nam olajšajo iskanje po besedilih, preštevanje pojavitev in računanje njihove statistične relevantnosti. Kot pri vseh jezikoslovnih dejavnostih pa tudi za terminografijo velja, da so zanjo korpusi z vsem pripadajočim instrumentarijem in raziskovalnimi metodami le vir surovih podatkov, naloga korpusnega terminografa pa je v prvi fazi izumljanje inteligentnih načinov dostopanja do podatkov in v drugi fazi njihova interpretacija. Teh faz namreč še nekaj časa ne bo mogoče avtomatizirati.

7.3 SAMODEJNO LUŠČENJE TERMINOLOGIJE

Samodejno luščenje terminologije (angl. *Automatic Term Extraction*) je področje računalniškega jezikoslovja, ki se ukvarja z metodami za računalniško prepoznavanje strokovnih izrazov v zbirkah strokovnih besedil. *Samodejno* navadno pomeni, da je naloga v celoti modelirana tako, da računalnik opravi ves postopek obdelave brez človeške pomoči ali posredovanja. Sistemi za luščenje terminologije so navadno res zasnovani tako, da se vhodni podatki, se pravi korpus, samodejno obdelajo, rezultat pa je seznam kandidatov terminov. Kljub temu je izraz nekoliko zavajajoč, saj gre s številnih vidikov za človeško nadzorovan proces. Na rezultate vplivamo že z izbiro korpusa, predvsem pa je v večini aplikacij predvideno, da avtomatsko pridobljene sezname terminov pred nadaljnjo uporabo pregleda terminolog ali strokovnjak. Nekatere metode so zasnovane tako, da uporabnik pred luščenjem sam določi parametre algoritma, na primer skladišne vzorce, ali pa luščenje poteka interaktivno, kjer uporabnik sproti preverja predlagane termine.

Razmerje med *avtomatskim* in *polavtomatskim* luščenjem še najlažje navežemo na namen uporabe. Kjer je razpoznavanje terminov le modul znotraj širše aplikacije, predvsem imamo v mislih sisteme za iskanje podatkov, poteka postopek povsem avtomatsko, rezultati obdelave pa se neposredno uporabijo pri iskanju dokumentov in njihovem razvrščanju. Za vse aplikacije, ki so namenjene terminološkemu delu in kjer je neposredni uporabnik človek, pa je verjetno primernejši izraz računalniško podprto, saj po luščenju neizogibno pride faza preverjanja in po potrebi čiščenja rezultatov.

V nadaljevanju poglavja pregledno predstavimo najpomembnejše metode pri avtomatskem luščenju terminologije. Čeprav pristope razdelimo na statistične, jezikoslovno usmerjene in hibridne, posvetimo pa se tudi nekaterim novejšim in naprednejšim tehnikam, je treba poudariti, da je takšna delitev zgolj poskus opredelitve usmerjenosti pristopa. Pri skoraj vseh gre v resnici za takšno ali drugačno kombinacijo jezikoslovnega znanja o naravi terminov in izrabo matematičnih lastnosti porazdelitve besed in besednih nizov v korpusih.

Če v grobem povzamemo težnje zadnjih desetih let, se v poznih osemdesetih in zgodnjih devetdesetih letih kaže pionirsko navdušenje nad statističnimi metodami. V tem času je korpusno jezikoslovje doživelo razmah, pogojen z nastankom prvih večjih korpusov, na primer BNC, ki so bili prvič prosto dostopni širši raziskovalni javnosti. Obenem so procesorske zmogljivosti osebnih računalnikov že dosegle raven, ko je bilo obdelave večjih količin podatkov mogoče opravljati brez Unixovih delovnih postaj, čeprav Unix in Linux v računalniškem jezikoslovju vse do danes ostajata razširjeni okolji.

Statistični pristopi tega časa so bili precej primitivni in so temeljili na pogostnostnih modelih jezika. Kmalu je postalo jasno, da je za boljše rezultate potrebno vključiti več jezikovnega znanja, pri čemer je najosnovnejša raven oblikoskladenjska analiza. V večini jezikov in strokovnih področij je ena od temeljnih značilnosti strokovnega jezika nominalizacija, zato se je pozornost usmerila na terminološko relevantne besedne vrste, se pravi najprej na samostalnike, nato pa na besedne zveze oziroma besednovrstne vzorce, ki so zajemali samostalniške zveze, takoj zatem pa še terminološko relevantne povezave *osebek* + *povedek* oziroma *povedek* + *predmet*.

To seveda zahteva vse globlje označevanje korpusov, v zadnjem primeru pa potrebujemo popolno stavčno analizo. Dokler je kazalo, da bo problem avtomatske skladdenjske analize vsak hip mogoče zadovoljivo rešiti, so si ti pristopi pomagali z mučno pridobljenimi ročno označenimi korpusi. A ker za mnoge jezike še danes ni na voljo

popolna skladišna analiza, je luščenje terminologije v zadnjih petih letih ponovno krenilo na bolj računalniško usmerjeno pot. Sodobne stohastične metode, kot je vektorsko modeliranje podobnosti jezikovnih enot, ki so se najprej razvile za namene iskanja podatkov in avtomatskega indeksiranja, so postale zanimive tudi za terminološke namene.

V času pisanja je ta trend še vedno opazen, pogojuje pa ga nagel razvoj računalniško-statističnih metod obdelave in modeliranja velikih količin podatkov. Strojno učenje in metaučenje sta danes uveljavljeni metodi za avtomatsko ugotavljanje pravil in vzorcev v velikih količinah podatkov na mnogih področjih, na primer v medicini, hidrometeorologiji, geologiji in geografiji, biokemiji in drugih. Čeprav je naravni jezik v smislu modeliranja eden najkompleksnejših sistemov, saj ni podrejen naravnim zakonom, se te metode kažejo kot zelo obetavne tudi v računalniškem jezikoslovju.

7.3.1 Statistični pristopi k iskanju terminov

Skupna lastnost statističnih pristopov je, da temeljijo na uporabi korpusov in izkoriščajo ponovljivost oziroma pogostost jezikovnih enot. Pri tem statistična obdelava služi preverjanju hipotez, ki jih imamo o lastnostih terminov v primerjavi z netermini, lastnostih strokovnih besedil v primerjavi z nestrokovnimi in o distribuciji strokovnega izrazja v različnih besedilnih vrstah in registrih.

Ena od glavnih predpostavk je, da so v strokovnih besedilih termini glavni nosilci pomena in tako predstavljajo »ključne« leksikalne enote. Ključnost smo predstavili že v razdelku o programu WordSmith Tools, na podoben način pa se uporablja tudi pri sistemih za samodejno luščenje terminologije.

Druga predpostavka, ki jo oblikujeta Justeson in Katz (1995), ugotavlja, da se »termini pojavljajo v izbruhih«. Če se v določenem delu strokovnega besedila leksikalna enota večkrat pojavi, je enota verjetneje termin. Merjenje zgolj korpusne frekvence torej ni dovolj.

Drugačno pojmovanje ključnosti se je razvilo v okviru iskanja podatkov, kjer skušamo v zbirki dokumentov najti tiste, ki najbolj ustrezajo poizvedbi. Tu naj bi dokumente, ki vsebujejo največ terminov iz poizvedbe, uvrstili na vrh seznama zadetkov, pri tem pa nam zgolj frekvenca termina v celotni zbirki dokumentov ne pomaga kaj dosti. Za merilo terminološkosti se torej uporablja tudi kriterij, v koliko dokumentih se iskana leksikalna enota pojavi.

Doslej smo previdno uporabljali izraz leksikalna enota – ta je namreč lahko beseda, še večkrat pa so termini večbesedne enote. Pri avtomatskem luščenju terminologije je zato razpoznavanje večbesednih enot izjemno pomembno, obenem pa se postavi nekaj vprašanj. Prvo je dolžina enote oziroma največje možno število besed v terminu, drugo problem gnezdenja, na primer *sistem polarnega žerjava* in v njem vsebovani *polarni žerjav*, tretje pa morebitna prekinjenost enot, na primer *sežigalnica ljubljanskih odpadkov*, kjer je termin le *sežigalnica odpadkov*.

Med prvimi poskusi luščenja terminologije s pomočjo korpusa in analize pogostosti je bil eksperiment Ahmada in soavtorjev (Ahmad in dr. 1992), v katerem so s pomočjo korpusnega orodja MATE²⁵ primerjali pogostosti pojavnic v strokovnem korpusu s področja mamografije s pogostostmi v izseku iz splošnega korpusa britanske angleščine LOB (Lancaster-Oslo-Bergen). Strokovni korpus je bil za današnje razmere smešno majhen, saj je obsegal le 28.458 pojavnic, izsek iz LOB-a pa je bil velik približno milijon pojavnic. Kljub temu avtorji pokažejo, da ob primerjavi relativne frekvence na površje splavajo prav tiste pojavnice, ki strokovni korpus najbolj označujejo.

Kljub temu so ugotovitve Ahmada in soavtorjev z današnjega vidika sporne, kajti kot poglobitni dokaz terminološkosti strokovnega korpusa navajajo dejstvo, da se med prvimi 50 najpogostejšimi pojavnicami v korpusu LOB ne pojavi niti ena beseda odprtega razreda oziroma produktivne besedne vrste, tj. samostalnik, pridevnik, glagol, prislov, medtem ko se na seznamu prvih 50 pojavnic strokovnega korpusa znajde kar 16 besed odprtega razreda, od tega 12 terminov. Primerjava ni relevantna, kajti velikosti obeh korpusov sta se tako zelo razlikovali, da je povsem umestno pričakovati, da se bo v enomilijonskem LOB-u s precejšnjo pogostostjo pojavilo več različnih oblik zaimkov, predlogov in števnikov kot v malem mamografskem korpusu.

7.3.1.1 TF-IDF

Na področju iskanja podatkov (angl. *Information Retrieval*) se je uveljavil model vektorskega prostora kot uspešna metoda za merjenje podobnosti med dokumenti in poizvedbo. Poenostavljeno povedano temelji ta model na dodeljevanju uteži²⁶ besedam v poizvedbi in v zbirki dokumentov, nato pa se na podlagi primerjave uteži posameznih besed v poizvedbi in v izboru dokumentov izvede razvrščanje dokumentov po pomembnosti (Manning in Schütze 1999: 543).

²⁵ <http://mate.nis.sdu.dk/>

²⁶ Utež je navadno številska vrednost, ki izraža pomembnost določene enote.

Dodeljevanje uteži posameznim enotam oziroma terminom – beseda termin se v kontekstu iskanja podatkov uporablja v pomenu ključna beseda – je pogojeno z dvema poglavitnima faktorjema: pogostostjo termina w_i v zbirki dokumentov d_j ($tf_{i,j}$) in številom dokumentov, v katerih se vsaj enkrat pojavi w_i . Oba faktorja povežemo v enotno utež, ki se imenuje *tf.idf* (angl. *Term Frequency – Inverse Document Frequency*). in se uporablja v več različicah z različnimi metodami normalizacije, v našem primeru v logaritmični obliki. Komponenta *idf* skrbi za to, da obtežimo termine, ki se pojavljajo le v redkih dokumentih, medtem ko je za besede, ki so prisotne v vseh dokumentih, vrednost *idf* enaka 0. Na ta način merimo specifičnost termina za določeno skupino besedil, izločimo pa t.i. prazne besede, ki so prisotne v vseh dokumentih.

Poudariti je treba, da se cenilka *tf.idf* izkaže za uporabno le, kadar imamo opravka s heterogeno zbirko dokumentov. Če bi s to metodo iskali termine v strokovnem korpusu določenega področja, bi našli le tiste enote, ki se pojavljajo v majhnem številu besedil, vse splošne termine področja pa bi zgrešili.

7.3.1.2 Drugi indikatorji terminološkosti

Strokovna besedila pa vsebujejo tudi termine, ki se ne pojavljajo pogosto. Celo nasprotno — zelo specifična poimenovanja z visoko terminološko vrednostjo se pogosto pojavijo le enkrat. Enopojavnice v korpusnem jezikoslovju imenujemo tudi *hapax legomena* in predstavljajo posebej zanimivo področje raziskovanja. Enopojavnice v velikih korpusih pogosto kažejo nove, ustvarjalne, sposojene ali napačno zapisane besede, se pravi elemente, ki bi jih sicer »iskali z lupo«, so pa vsekakor vredni jezikoslovne pozornosti.

V okviru strokovnih besedil med enopojavnicami srečamo manj literarnih novotvorjenk, zato pa precej tujk, sposojenk in za stroko specifičnih poimenovanj. Žal je med njimi tudi precej napačno črkovanih besed, zato je avtomatska obdelava takih seznamov nekoliko otežena.

Podobno merilo »izjemnosti« je tudi, če je pojavnica neznana lematizatorju ali črkovalniku. Ti programi so navadno opremljeni s precej velikimi leksikoni besed in odsotnost besede je lahko znak nesplošnosti, se pravi potencialne terminološkosti pojavnice.

Poleg besed, ki na tak ali drugačen način izstopajo s svojo redkostjo ali neznano-stjo, pa je še nekaj drugih indikatorjev terminološkosti. Že med neznanimi besedami

opazimo precej imen, krajšav in akronimov. Terminološka vrednost imen je odvisna od področja in tipa besedila, a v mnogih primerih so imena in eponimske izrazje pomemben del terminologije stroke. V jedrskem korpusu, ki je bil uporabljen za raziskavo s področja luščenja izrazja (Vintar 2003), najdemo eponimske izraze, kot so *Jedrska elektrarna Krško*, *Petrijeve mreže*, *Siemens Framatome*, *Ostwald*, *Westinghouse*, ki bi jih za prevajalsko-terminološke namene verjetno uvrstili med kandidate. V slovenščini je izimensko izrazje včasih težko prepoznati zaradi pisave z malo začetnico, še vedno pa mnogo lažje kot na primer v nemščini, kjer se z veliko začetnico piše vse samostalnike.

Še posebej pomemben tip izrazja predstavljajo akronimi, ki jih je v nekaterih vrstah strokovnih besedil toliko, da je njihovo tolmačenje za nestrokovnjaka skoraj nemogoče. Ker je avtomatsko iskanje akronimov razmeroma enostavno, v njihovi bližini pa pogosto najdemo tudi razvezano obliko, so koristni tudi pri pomenski analizi besedila.

Pokazatelj terminološkosti pa so lahko tudi sorodnice. Ker so strokovni jeziki še bolj odprti do prevzemanja izrazja kot splošni jezik, je v njih mnogo besed s tuji koreni, ali pa so prevzete povsem citatno. S primerjavo nizov črk je sorodnice mogoče avtomatsko izluščiti, tuj izvor besede pa včasih pomeni terminološko relevantnost.

7.3.1.3 Statistike za luščenje kolokacij

Termini so v veliki večini večbesedne enote, ki so glede svoje oblike razmeroma strogo določene. S tega vidika torej termini predstavljajo stalne besedne zveze, zato ni presenetljivo, da se mnogi statistični pristopi k luščenju terminologije problema lotijo s predpostavko, da so termini pravzaprav vrsta kolokacij.

Za iskanje kolokacij se je v korpusnem jezikoslovju razvila že vrsta metod, pregledno predstavitev najdemo v Manning in Schütze (2000: 151-191). Prvi indikator, da določena besedna zveza predstavlja kolokacijo, je že njena absolutna pogostost. Za terminološke namene je koristno, da iz takšnega seznama izločimo prazne besede, ki se pojavljajo na začetku ali na koncu besedne zveze. Tabela 8 in Tabela 9 prikazujeta primer takšnih izluščenih enot iz vojaškega korpusa Grizold.

Tabela 8: Dvobesedne enote po absolutni pogostosti

w1	w2	f(w1w2)	f(w1)	f(w2)
ki	so	3070	16293	25239
se	je	2778	14567	40293
ki	je	2644	16293	40293
da	je	2177	12893	40293
pa	so	2056	10301	25239
pa	je	2055	10301	40293
je	bil	1942	40292	2458
naj	bi	1915	2306	8204
je	bila	1867	40292	2382
da	bi	1643	12893	8204
je	bilo	1620	40292	2477
so	se	1594	25239	14567
da	so	1549	12893	25239
je	v	1440	40292	27862
tako	da	1347	4925	12893
pa	je	1294	6308	40293
da	se	1282	12893	14567
ki	se	1127	16293	14567
so	v	936	25239	27862
so	bili	852	25239	1137
kar	je	774	2641	40293
ki	bi	749	16293	8204
pa	se	727	10301	14567
so	ga	712	25239	3007
ki	ga	711	16293	3007
bi	se	709	8204	14567
so	jih	667	25239	2709
ki	jih	658	16293	2709
ali	pa	651	5042	6308

Tabela 9: Dvobesedne enote po absolutni pogostosti z izločanjem praznih besed

w1w2	f(w1)	f(w2)
še vedno	853	434
svetovne vojne	918	286
svetovni vojni	948	282
oboroženih sil	1239	269
še posebej	523	223
drugi svetovni	299	217
svetovno vojno	353	188
drugi strani	722	168
druge svetovne	323	166
potem ko	2105	166
balističnih raket	1380	163
letalskih sil	1239	144
kratkega dosega	462	142
zračne obrambe	559	130
bojne glave	193	127
zaradi česar	193	126
vse bolj	1150	125
še bolj	1150	124
mornariške pehote	260	122
proti cilju	171	121
srednjega dosega	462	118
še naprej	504	117
vojne mornarice	345	116
sovjetska zveza	146	114
drugo svetovno	196	112
sovjetski zvezi	267	112
operativno uporabo	861	111
prvi vrsti	191	106
vojna mornarica	325	106
smodniških plinov	235	104

Očitno je, da zgolj pogostost ni dovolj za odločanje o tem, kaj je kolokacija in kaj ne, čeprav je rezultat po izločanju praznih besed že bistveno boljši. V korpusnem jezikoslovju se za luščenje kolokacij uporabljajo različne statistike, pri čemer ima skoraj vsaka svoje dobre in slabe lastnosti. Vse temeljijo na primerjavi opaženih in pričakovanih pogostosti, s čimer se potrjuje ali ovrže hipotezo o neodvisnosti spremenljivk, v našem primeru posameznih besed.

Ena pogosto omenjanih statistik, ki se uporablja za odkrivanje kolokacij, je Pearsonov test χ^2 (Manning in Schütze 2000: 169). Če želimo preveriti, ali določena v korpusu opažena besedna dvojica predstavlja kolokacijo, seštejemo vse razlike med opaženimi in pričakovanimi pogostostmi (O_{ij} in E_{ij}), kjer je O_{ij} opažena pogostost besedne dvojice, E_{ij} pa je pričakovana pogostost, ki jo izračunamo iz posameznih pogostosti obeh opazovanih besed deljeno z N , ki predstavlja velikost korpusa.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Statistika χ^2 postane nezanesljiva, če imamo opravka z majhnimi pogostostmi. Skozi desetletja razvoja računalniškega jezikoslovja se je celo bolje izkazala pri drugih nalogah, denimo za iskanje prevodnih ustreznic v vzporednem korpusu ali kot merilo podobnosti med korpusi (Kilgariff 2001).

Za neterminološke namene se uporablja še statistika vzajemnosti (*Mutual Information*) ali njena izpeljanka MI^3 . Pri terminološki rabi se tudi tu pojavi problem, da MI na vrh seznama kolokacijskih kandidatov postavi redko opažene besedne dvojke.

Tabela 10: Primeri dvobesednih enot, izluščenih z vzajemnostjo (MI)

w1	w2	rank	MI	f(w1w2)	f(w1)	f(w2)
dokopati	rešilne	1	20.4847	1	1	1
manjšajo	enakopravneje	1	20.4847	1	1	1
Iraški	lanserji	1	20.4847	1	3	6
pričujočemu	testu	1	20.4847	1	1	2
goalkeeper	Osnovo	1	20.4847	1	1	5
Vizualni	dražljaji	1	20.4847	1	1	1
NASPROTUJEJO	Propaganda	1	20.4847	1	1	1
referencami	Hughesu/Douglasu	1	20.4847	1	1	1
gusarjev	Hajrudina	1	20.4847	1	1	1
perjem	U-2/TR-1	1	20.4847	1	1	1
grabijo	podlage	1	20.4847	1	1	1
seeks	US	1	20.4847	1	1	5
TOTC	Time	1	20.4847	1	1	1
strankinim	Veljal	1	20.4847	1	1	1
nezahtevni	10-sedežni	1	20.4847	1	1	1
izpitno	Proga	1	20.4847	1	1	1
decembrskega	preformiranja	1	20.4847	1	1	1
urejenih	cestišč	1	20.4847	1	1	1
hrbtenico	Pretres	1	20.4847	1	8	1
Watervliet	Arsenalu	1	20.4847	1	1	1
jekleno-aluminijevega	kompozita	1	20.4847	1	1	1
hlapljivi	komponenti	1	20.4847	1	1	2
Texture	Mats	1	20.4847	1	1	1
NEW	YORK	1	20.4847	1	1	1
Porschejevih	izboklin	1	20.4847	1	1	1
FM	USB	1	20.4847	1	1	1
komplicirani	Pomaga	1	20.4847	1	1	1
arzenali	Dolgo	1	20.4847	1	1	2
Kurt	Buhlingen	1	20.4847	1	1	1

Daille (1995) primerja vrsto različnih statističnih metod z ročno označenimi leksikalnimi enotami in ugotavlja, da se statistika logaritma razmerij verjetja (log-likelihood ratio) še najbolj ujema z človekovim pojmovanjem večbesednih enot. Dejansko se tudi za terminološke namene ta izkaže za najbolj obetavno.

$$\begin{aligned} \log\text{--}likelihood &= a \log a + b \log b + c \log c + d \log d \\ &- (a + b) \log(a + b) - (a + c) \log(a + c) \\ &- (b + d) \log(b + d) - (c + d) \log(c + d) \\ &+ (a + b + c + d) \log(a + b + c + d) \end{aligned}$$

Vrednost LL se računa po zgornji enačbi, kjer so a, b, c in d pogostosti besedne enote (i, j) v kontingenčni matriki.

	i	≠i
j	a	b
≠j	c	d

Danes metode, ki se opirajo izključno na statistiko besednih oblik, za luščenje terminologije niso več najbolj uporabne. Za mnogo svetovnih jezikov imamo namreč na voljo oblikoslovne označevalnike, s katerimi lahko besednim oblikam pripišemo leme in oblikoslovne oznake, te informacije pa nato bistveno vplivajo na metode, in posledično rezultate, samodejnega luščenja.

7.3.2 Luščenje terminov s pomočjo oblikoskladenjskih vzorcev

Termini so pogosto samostalniške besedne zveze, te pa sledijo predvidljivim oblikoskladenjskim vzorcem. Z luščenjem tipičnih vzorcev tako izluščimo tudi termine. Med prvimi to sta to metodo predlagala Dagan in Church (1994), takšno ali drugačno različico pa opisujejo še Justeson in Katz (1995), Bourigault (1996), Heid in dr. (2001), Jacquemin (2001), Vintar (2002) in drugi. Tipične samostalniške vzorce v slovenščini smo opisali že v Tabela 3, izbor vzorcev za namene luščenja terminologije pa je odvisen od namena luščenja, strokovnega področja in značilnosti korpusa. Izluščene besedne zveze še niso termini, saj denimo vzorcu P + S lahko ustrezata zvezi *matična plošča* in *naslednji primer*. Za razvrščanje besednih zvez po terminološki relevantnosti se uporablja vsota ključnosti oziroma relativnih pogostosti posameznih besed. Tabela 11 kaže prvih štirideset terminoloških kandidatov, izluščenih z metodo vzorcev in terminoloških uteži, iz vojaškega korpusa Grizold.

Tabela 11: Terminološki kandidati, izluščeni s pomočjo oblikoskladenjskih vzorcev

Verjetno je še največja prednost metode z luščenjem vzorcev, da ni odvisna od pogostosti posameznih leksikalnih enot, poleg tega pa je pri tej metodi mogoče mnogo bolj učinkovito obravnavati problem gnezdenih terminov.

1	oborožena sila	21	zalivska vojna
2	ognjena podpora	22	vojaško izobraževanje
3	ognjena moč	23	poveljstvo bataljona
4	ognjeni položaj	24	ognjeni sistem
5	konec leta	25	sistem zračne obrambe
6	pomembnejša vloga	26	poveljniška četa
7	usposabljanje vojakov	27	poveljnik voda
8	oborožen boj	28	štabni proces
9	krizno upravljanje	29	naslednje leto
10	pripadnik stalne sestave	30	oblika usposabljanja
11	artilerijski bataljon	31	nadzor zračnega prostora
12	poveljnik bataljona	32	nadaljnji razvoj
13	osamosvojitvena vojna	33	top kalibra
14	pehotni oddelek	34	višja stopnja
15	načelnik generalštaba	35	vodni podčastnik
16	vojna leta	36	spopad nizke intenzivnosti
17	večje število	37	usposabljanje kandidatov
18	organizacijska enota	38	pripadnik oboroženih sil
19	obrambno načrtovanje	39	vojaški kaplan
20	zvezna vojska	40	poveljstvo brigade

7.3.3 Iskanje prevodnih ustreznic

Besedna poravnava (*Word Alignment*) je izraz za tehnologijo statističnega pridobivanja leksikonov prevodnih ustreznic iz vzporednih korpusov. Metoda temelji na predpostavki, da je za vsako besedo v korpusu iz prevodnih segmentov, kjer se pojavlja, mogoče izračunati najverjetnejšo ustreznico.

Danes je za avtomatsko izdelavo dvojezičnih leksikonov na voljo več prosto dostopnih programov, npr. Giza++ (Och in Ney 2003), UPlug (Tiedemann 1999) in

Twente (Hiemstra 1998). Algoritmi in njihove nastavitve se med seboj sicer precej razlikujejo, vsem skupna lastnost pa je, da program za vsako besedo v korpusu predlaga eno ali več ustreznih skupaj z verjetnostmi prevoda. Rezultate besedne poravnave lahko izboljšamo, če iz korpusa prej odstranimo prazne besede, in če za poravnavo uporabljamo lematiziran korpus.

Sorodnice (*cognates*) so besede, ki so si po zunanji podobi, navadno pa tudi po pomenu, podobne, npr. *sistem/system*, *informacija/information*, *Francija/France*. Med njimi najdemo precej imen in kratic, večinoma pa gre za besede neoklasičnega izvora, ki so si v številnih jezikih podobne. V jezikovnih tehnologijah se sorodnice izkorišča predvsem za izboljšanje stavčne poravnave (Simard in dr. 1992), zanimive pa so tudi s prevodoslovnega vidika (Vintar in Hansen 2002).

Ko smo izluščili termine obeh jezikov in presejali sezname kandidatov, se znajdemo pred naslednjo nalogo — poiskati pare prevodnih ustreznih. Pri enobesednih terminih so ustreznice seveda vključene že v dvojezični leksikon, problem predstavljajo le primeri, kjer se enobesedni termin prevede v večbesednega ali obratno (npr. *uparjalnik/steam generator*). Če je večbesedni termin povsem ustaljen, je ta pojav mogoče ugotoviti na ravni besedne poravnave, saj dobimo za *uparjalnik* ustreznici *steam* in *generator*, vsako z verjetnostjo 0,50. Ta metoda je zanesljiva le pri pogostih terminih in le, če smo za izdelavo dvojezičnega leksikona uporabili lematizirani korpus.

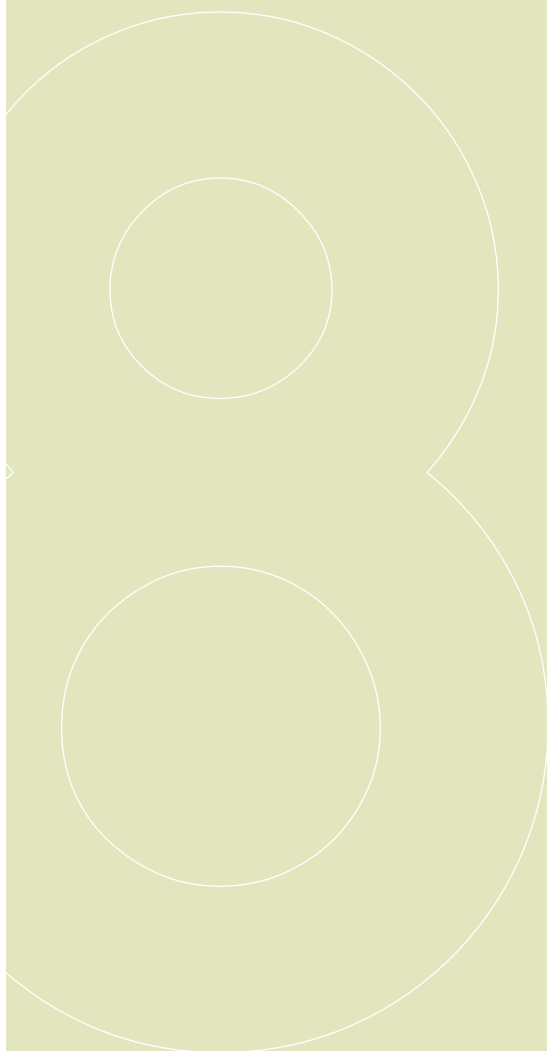
V vseh ostalih primerih pa skušamo ustreznico večbesednega termina najti tako, da za vsako od besed v izvirnem terminu v leksikonu poiščemo ustreznice, nato pa med ciljnim terminom najdemo tistega, ki vsebuje največ ustreznih. Zanesljivost prevoda izrazimo s seštevkom prevodne ustreznosti vsake posamezne ustreznice.

projekt	zamenjave	uparjalnikov
↓	↓	↓
project 1.00	[null] 1.00	steam 0.49 generator 0.33 generators 0.18

Tu med angleškimi termini največ ustreznih vsebuje izraz *steam generator replacement project*, ki ga izberemo s prevodno verjetnostjo 1,82.

Ta način omogoča dokaj precizno ugotavljanje parov ustreznih, njegova prednost pa je, da razmeroma dobro deluje tudi z nelematiziranimi termini. Kljub temu opazimo, da za številne kandidate v enem jeziku ne najdemo ustreznih, zato je skupno število dvojezičnih parov manjše od števila najdenih terminov v posameznih jezikih.

8 Ustvarjanje terminoloških baz



Zielinski in Ramirez (2005) v svoji raziskavi ugotavljata, da kar 68,7 odstotkov poklicnih prevajalcev za upravljanje terminologije uporablja specializirano tovrstno orodje, ob tem jih okrog 50 odstotkov občasno za hrambo terminoloških podatkov uporablja tudi Excel ali Word, le 5,4 odstotki pa se zanašajo na sistem listkovne kartoteke. Prednosti specializiranih terminoloških orodij so številne, saj omogočajo individualizirano strukturiranje terminološke baze glede na specifične potrebe projekta, ponujajo napredne možnosti prikaza, razvrščanja, sejanja in urejanja vnosov, prek standardiziranih načinov uvoza in izvoza podatkov pa podpirajo tudi migracijo terminoloških podatkov med različnimi aplikacijami in sistemi.

Ker terminološka orodja zvečine razvijajo proizvajalci prevajalskih tehnologij, so vsi razširjeni programi neposredno združljivi s prevajalskim namizjem istega proizvajalca.

Nekaj vodilnih izdelkov na področju upravljanja terminologije je v času pisanja naslednjih:

- STAR Webterm in Termstar²⁷,
- SDL MultiTerm²⁸,
- Lingo²⁹,
- Cycom³⁰ (prosto dostopno orodje s podporo za standard TBX).

Ker je od naštetih daleč najbolj razširjeno orodje SDL MultiTerm, v nadaljevanju sledi opis izdelave terminološke baze v tem programu. To seveda ne pomeni, da je MultiTerm edina možna izbira za ustvarjanje terminološke baze, saj izkušnje uporabnikov z MultiTermom pri gradnji velikih terminoloških zbirk kažejo tudi na precej pomanjkljivosti.

Na Terminološki komisiji ZRC SAZU, kjer še vedno nastajajo najobsežnejša terminografska dela, so se odločili za razvoj lastne aplikacije z imenom SlovarRed (Košmrlj-Levačič in Seliškar 2004), ki temelji na programu MS Access in dobro služi potrebam tamkajšnjih terminografov.

²⁷ http://www.star-ts.com/STAR/eng/products5_webterm.htm

²⁸ <http://www.sdl.com/en/products/products-index/multiTerm.asp>

²⁹ <http://www.lexicoool.com/soft.asp>

³⁰ <http://club.cycom.co.uk/termManage.html>

8.1 USTVARJANJE TERMINOLOŠKIH BAZ S PROGRAMOM SDL TRADOS MULTITERM

SDL Trados MultiTerm je terminološko orodje, ki se je zaradi svoje prilagodljivosti uporabniškim potrebam in navezave na prevajalska namizja že v devetdesetih letih razmahnilo med prevajalci in terminologi. Danes ga uporablja na desettisoče uporabnikov po vsem svetu, ne le za prevajalske potrebe, ampak tudi za resnejše terminografske projekte. Poseben premik na tem področju je bil dosežen s pojavom strežniške različice, ki skupinam uporabnikov omogoča uporabo in gradnjo terminološke baze na daljavo.

Poglavitna prednost programa MultiTerm je, da program s svojo zasnovo omogoča oblikovanje poljubnih podatkovnih kategorij, obenem pa sugerira njihovo strukturiranje po pojmovnem načinu. Terminološke baze so lahko mnogojezične, program prek Unicoda podpira različne znakovne nabore, vse baze pa so samodejno obrnljive.

MultiTerm omogoča več načinov iskanja po bazi:

- enostavno iskanje po besedi ali frazi (drsní ležaj),
drsní ležaj
- iskanje z nadomestnimi znaki (*ležaj*),
aksialni kotalni ležaj, dinamična nosilnost ležaja, prilagodni kroglični ležaj
- megleno iskanje, ki poišče tudi podobne besede (~dinamic),
basic dynamic load rating

Za ustvarjanje nove terminološke baze v MultiTermu lahko uporabimo eno od glosarskih predlog, ki jih ponuja program, za naprednejše uporabnike pa je bolj smiselno, da si strukturo baze definirajo sami. Pri tem je predvsem pomembno, da pred ustvarjanjem nove baze dobro razmislimo o podatkovnih kategorijah, ki jih nameravamo uporabiti, o povezavah med njimi in o celotni strukturi terminološkega vnosa (glej razdelek 5.2). Ko baza namreč že vsebuje nekaj vnosov, je strukturo razmeroma zahtevno spreminjati — nove podatkovne kategorije sicer lahko brez težav dodajamo, spreminjanja že zapolnjenih polj pa program ne dopušča.

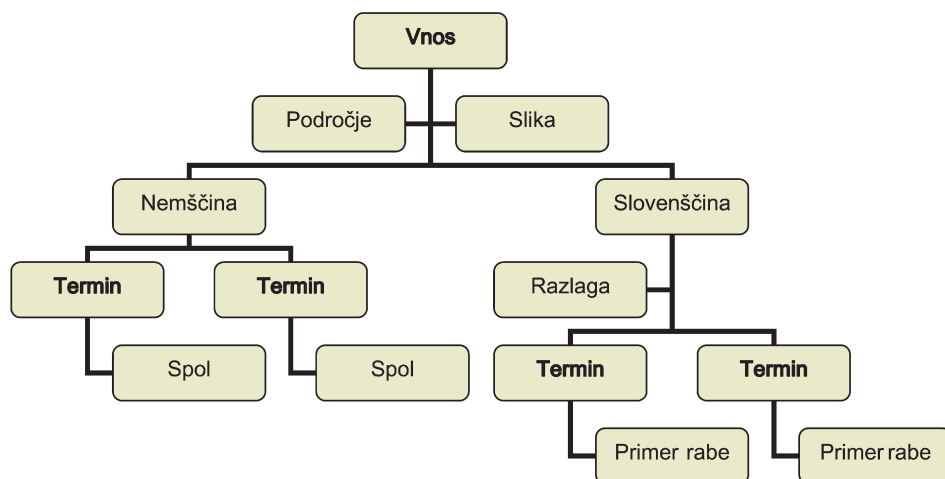
Za primer vzemimo, da ustvarjamo dvojezični nemško-slovenski glosar puškarstva.³¹ Glosar je namenjen slovenskim uporabnikom, zato bodo predstavljeni puškarski pojmi vsebovali razlago v slovenskem jeziku, poleg slovenskih in nemških poimenovanj pa bodo vključeni še primer rabe v slovenščini, oznaka spola za nemške termine,

³¹ Primeri so vzeti iz glosarja puškarstva, ki ga je v okviru predmeta Prevajalska orodja v študijskem letu 2007/2008 izdelala Vida Frelih.

ponekod pa tudi slika, saj glosar vsebuje precej poimenovanj delov pušk, ki jih je brez slikovnega gradiva težko opisati. Glosarski vnosi bodo razdeljeni na podpodročja puškarstva.

Ko smo opredelili podatkovne kategorije, ki jih bo glosar vseboval, moramo zasnovati še njihovo strukturo. Primer možne strukture prikazuje Slika 7.

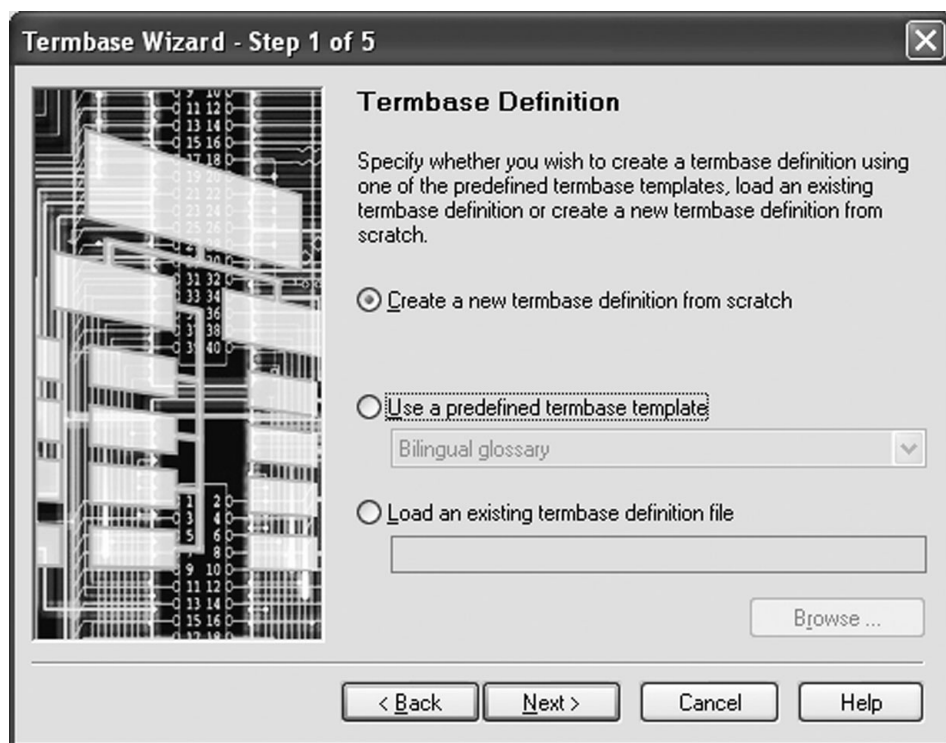
Slika 16: Primer strukture terminološkega vnosa



Razmisлити moramo tudi o tem, katere podatkovne strukture so v bazi obvezne in jih mora vsebovati vsak vnos, katere so obvezne in katere poljubne, pa tudi kakšno vsebino bodo imele posamezne kategorije. Če želimo podatke v bazi čimbolj poenotiti, lahko za vse podatkovne kategorije z vnaprej določljivo vsebino določimo izbirni seznam (npr. Spol: m, ž, s).

Zdaj lahko v programu MultiTerm ustvarimo novo bazo, in sicer z ukazi **Termbase – Create Termbase....** Ko smo določili, v katero mapo naj program shranjuje datoteke nove baze, vstopimo v čarovnika, ki nas vodi po posameznih korakih ustvarjanja baze. V prvem koraku imamo na voljo tri načine ustvarjanja baze (Slika 17).

Slika 17: Ustvarjanje baze v MultiTermu — prvi korak



Prva možnost je ustvarjanje lastne baze s strukturo vred, se pravi brez uporabe predlog, ki so priložene programu. Druga možnost vključuje uporabo ene od predlog, tretja pa predvideva, da želimo uporabiti strukturo baze, ki smo jo ustvarili že kdaj prej. V našem primeru izberemo prvo možnost.

V drugem koraku bazo poimenujemo. Ime, ki ga izberemo tu, bo MultiTerm uporabljal le za prikaz baze v okviru programa. Vse datoteke, ki jih program samodejno ustvari ob novi bazi, se bodo imenovala drugače. V tretjem koraku določimo jezike, ki jih bo baza vsebovala, pri čemer lahko jezikom po želji pripišemo slovenska imena. Program bo za vsak jezik samodejno predvidel poljubno število polj za vnos termina.

V četrtem koraku definiramo vse ostale podatkovne kategorije, se pravi Področje, Slika, Razlaga, Primer rabe in Spol. Privzeti tip vseh teh podatkovnih kategorij je besedilo (*Text*), kar pa zlahka spremenimo s klikom na tipko **Advanced...** Odpre se okence,

kjer lahko izbiramo med različnimi tipi podatkovnih polj, za izbirna polja (*Picklist*) pa lahko tudi določimo možne vrednosti (Slika 18). Slikovno gradivo, zvočni ali video posnetek se lahko vnašajo v polje, ki je definirano kot večpredstavnostno (*Multimedia*).

Slika 18: Določanje tipa podatkovne kategorije



V naslednjem koraku določimo še strukturo baze. MultiTerm ima vnaprej določene tri ravni, na katere se uvrščajo podatkovne kategorije, in sicer raven pojma oziroma vnosa (*Entry level*), raven jezika (*Index level*) in raven termina (*Term level*). V skladu s prej opisano strukturo baze bomo uvrstili kategoriji Slika in Področje na raven vnosa, kategorijo Razlaga na raven jezika in kategoriji Spol in Primer rabe na raven termina. Ko je struktura takšna, kot si jo želimo, zaključimo čarovnika in v programskem oknu se prikaže modra pločevinka z imenom glosarja, kar pomeni, da je baza pripravljena na vnašanje podatkov.

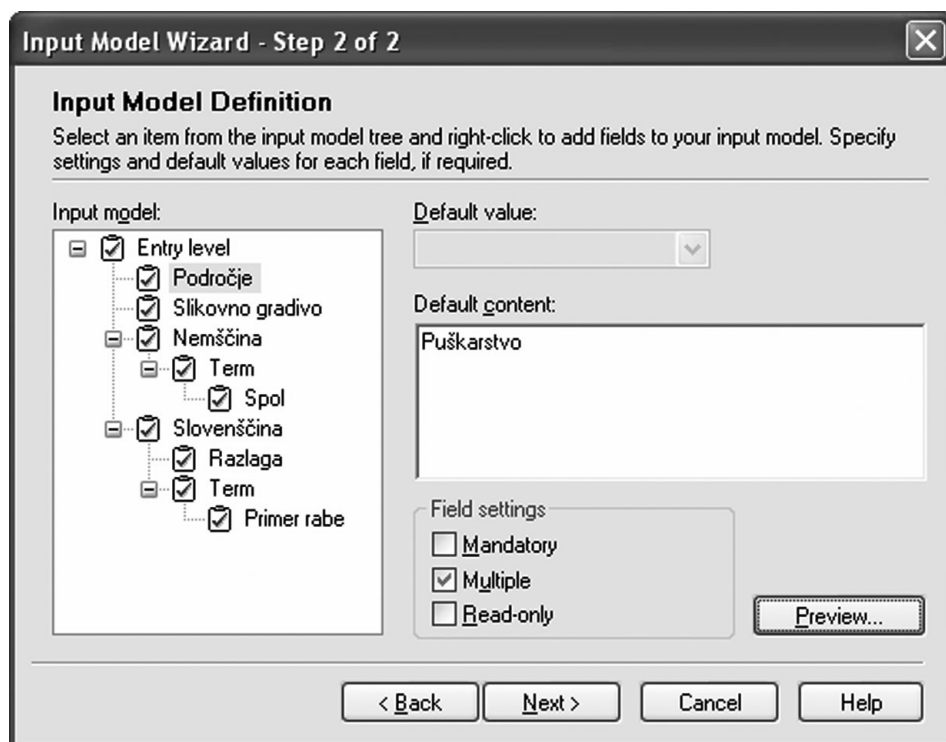
Za vnašanje podatkov imamo na voljo dva načina. Prvi je, da z uporabo menijskega ukaza **Entry – Add** oziroma bližnjice **F3** preklopimo v način za vnašanje, kjer se programsko okno obarva sivo in se prikažejo osnovne vnosne ravni za vse jezike baze. S klikom na posamezno polje se odpre okence, kamor lahko vtipkamo besedilo. Za vnašanje besedilnih in drugih dodatnih polj z desnim gumbom miške kliknemo na ime posamezne ravni in prikaže se meni, ki vsebuje vse razpoložljive podatkovne

kategorije na tisti ravni. Tako v primeru naše baze z desnim klikom na *Entry* dobimo meni s kategorijama Slika in Področje, z desnim klikom na Nemščina dobimo kategoriji Term in Razlaga, z desnim klikom na posamezni termin pa lahko odpremo polji Spol in Primer rabe.

Ko smo pod en terminološki vnos dodali vse podatke, ga shranimo s tipko **F10** ali menijskim ukazom **Entry – Save** ali desnim klikom in izbiro **Save**.

Drugi način za vnašanje je s pomočjo vnosne maske (*Input Model*), ki si jo definiramo v meniju **Entry — Input Models....** Najprej se odpre okence, kjer je na voljo privzeti način za vnašanje Default input model. Če želimo ustvariti lastno masko, kliknemo **Create...** in zažene se čarovnik, ki nas vodi skozi postopek. V prvem koraku določimo ime maske, v drugem pa znova natančno strukturo in zaporednje vnosnih polj. Tu lahko določimo tudi, ali je polje obvezno (Mandatory) in ali se lahko pojavi znotraj enega vnosa več kot enkrat (Multiple). Če ima polje privzeto vsebino, ki se naj ne bi spreminjala, ga označimo kot samo za branje (Read-only).

Slika 19: Ustvarjanje lastne maske za vnašanje



Vnosna maska, ki jo ustvarimo po meri, zagotavlja enovitost glosarske baze, saj bodo posledično vsi vnosi opremljeni na enak način. To seveda ne pomeni, da morajo biti pri vseh vnosih vsa polja izpolnjena — če k določenemu vnosu ne želimo pripisati razlage, imamo vso svobodo, da tega ne storimo, če le polja Razlaga nismo označili kot obveznega. Pri obveznih poljih pa program zahteva, da so izpolnjena, sicer vnosa ne moremo shraniti.

V tako pripravljeno okolje lahko zdaj vnašamo terminološke podatke. Če med izdelavo terminološke baze ugotovimo, da je struktura nepopolna in da želimo denimo dodati novo podatkovno kategorijo, to storimo na enostaven način z ukazom **Termbase - Modify Termbase Definition....** Spreminjanje tistih kategorij, ki so že zapolnjene s podatki, pa na ta način ni več mogoče. V primeru, da želimo podatkovne kategorije zgolj preimenovati, lahko bazo izvozimo v obliko XML in spremembe opravimo s funkcijo Zamenjaj v enem od urejevalnikov besedil. Za spremenjeno bazo moramo nato na novo definirati podatkovne kategorije, nato lahko bazo spet uvozimo v program MultiTerm.

MultiTerm podpira različne oblike izvoza terminološke baze. Privzeta oblika, v katero lahko izvozimo podatke, je XML, pri čemer program uporablja zapis, ki je podoben standardu TBX, ni pa z njim skladen (Slika 20). Druge oblike izvoza so še HTML, zapis MultiTerm5, RTF in golo besedilo, kjer so podatkovni stolpci ločeni s tabulatorji. Izvozne formate je mogoče tudi spremeniti in prilagoditi.

Slika 20: Izvoz iz programa MultiTerm v XML

```
<mtf>
<conceptGrp>
  <concept>1</concept>
  <transacGrp>
    <transac type="origination">NN</transac>
    <date>2007-12-26T12:47:17</date>
  </transacGrp>
  <transacGrp>
    <transac type="modification">NN</transac>
    <date>2007-12-26T12:47:17</date>
  </transacGrp>
  <languageGrp>
    <language type="English" lang="EN"/>
    <termGrp>
      <term>offensive rebound</term>
      <transacGrp>
        <transac type="origination">NN</transac>
        <date>2007-12-26T12:47:17</date>
      </transacGrp>
      <transacGrp>
        <transac type="modification">NN</transac>
        <date>2007-12-26T12:47:17</date>
      </transacGrp>
      <descripGrp>
        <descrip type="Context">Geehan was again in the middle of the action
as she pulled down the offensive rebound and fed Wrice for an open three.</
descrip>
      </descripGrp>
    </termGrp>
```

Pretvarjanje obstoječih glosarjev v MultiTermovo obliko je bilo некоč kompleksno in zamudno opravilo, od različice 8 naprej pa MultiTerm neposredno podpira uvoz iz Excelove tabele.

8.2 POGOSTE NAPAKE

O izdelavi terminoloških glosarjev smo doslej povedali že kar nekaj, za resno terminografsko delo pa so zagotovo potrebne tudi ustrezne izkušnje. V tem razdelku navajamo nekaj pogostih napak, ki se pojavljajo pri gradnji terminoloških baz predvsem v kontekstu poučevanja terminografije študentom prevajalstva.

Slika 21 kaže *farming* vnos iz terminološke baze s področja ekološkega kmetovanja, ki so ga izdelale študentke 3. letnika Oddelka za prevajalstvo FF. Gre za zagotovo osrednji pojem, ki pa se na področju ekološkega kmetovanja pojavlja predvsem kot glavna sestavina terminoloških samostalniških besednih zvez za poimenovanje različnih načinov kmetovanja. Predstavljeni vnos je torej primer za ponesrečeni poskus prikaza pojmovnega polja s podpomenkami vred, saj uvaja polje Hyponym, to pa vsebuje izraze, ki prav tako označujejo osrednje pojme in si torej zaslužijo lastne vnose. Poleg tega sta *conventional farming* in *traditional farming* vzporedni poimenovanji za isti pojem, kar iz vnosa ni razvidno. Tudi definicija ni oblikovana v skladu s smernicami, saj se prične s slovnično analizo namesto z najbližjo nadpomenko. V nadaljevanju pa definicija omenja še dve alternativni poimenovanji za isti pojem v slovenščini, in sicer kmetijstvo in agrikultura. Če gre res za vzporedni poimenovanji, bi ju morali vnesti kot ločena termina.

Slika 21: Primer iz študentske terminološke baze o ekološkem kmetovanju

Entry number: 50	
	English
Term: farming	
	Hyponym: organic farming
	Hyponym: conventional farming
	Hyponym: traditional farming
	Slovenian
Term: kmetovanje	
	Hyponym: ekološko kmetovanje
	Hyponym: konvencionalno kmetovanje
	Hyponym: tradicionalno kmetovanje
	Definition: glagolnik od kmetovati, tudi kmetijstvo. Kmetijstvo ali agrikultura je ena osnovnih in prvotnih človekovih dejavnosti. Delimo jo na poljedelstvo in živinorejo. Znanstvena veda o kmetijstvu je agronomija.

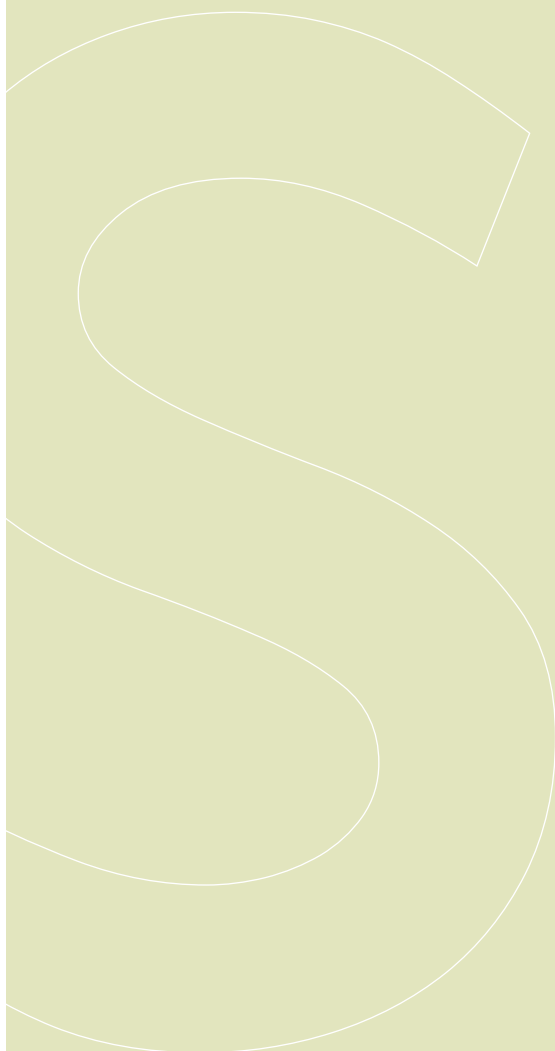
Slika 22: Primer vnosa iz študentske terminološke baze o vinarstvu

Entry number: 16
 English
Term: certified geographic origin
Definition: is based on the largest viticultural territory unit, the viticultural region. This is the most permissive of the standard declarations of origin. The grapes for PGP wines must be grown within one viticultural region (that is, Podravje, Posavje or Primorje), although the wine may be fermented and bottled outside the region. The label applies to "dezelno vino" ("regional wine") which is classified as table wine.
Context: We serve excellent opened wines with a certified geographic origin.
Field: besede, povezane z vinarstvom
 Slovenian
Term: priznano geografsko poreklo (PGP)
Context: Nudimo odlična odprta vina s priznanim geografskim poreklom.
Field: besede, povezane z vinom

Na Slika 22 je primer iz baze o vinarstvu, ki že na ravni termina krši načelo elementarnosti. Polje Term v slovenščini namreč vsebuje dve poimenovanji, in sicer izpisani termin *priznano geografsko poreklo* in krajšavo *PGP*. To je pri neukih uporabnikih MultiTerma zelo pogosta napaka, ki se pojavlja v različnih oblikah, na primer z vključevanjem določnega člena v indeksno polje (angl. *the Big Bang* [astronomija], nem. *das Norovirus* [medicina]), z zapisovanjem terminoloških variacij z oklepajem (sl. (*splet*na) *klepetalnica* [informatika]), z naštevanjem sopomenk v istem polju, ločenih z vejico (sl. *okužba, infekt* [medicina]) itd.

Pogosta strukturna napaka, ki se pojavlja predvsem, kadar za definicijo baze uporabimo eno od MultiTermovih predlog, je uvrščanje definicije na raven termina. To napako izkazuje tudi baza, iz katere je vzet primer na Slika 21. Kot smo pojasnili že nekajkrat, sodita definicija oziroma razlaga na raven jezika, saj se tipično nanašata na vsa alternativna poimenovanja pojma.

Stvarno kazalo



B

besednovrstni vzorec 40, 100
BNC 100

C

citatni prevzem 52, 104

D

diskurz 13, 14, 17, 20, 37, 38, 46, 48
dokumentalist 9, 47
dolžina termina 40

E

eponim 53, 54, 104

F

FIDAprus 30, 31, 41, 52, 80

G

gospodarnost 47, 53, 56

H

hapax legomena 103
homogenost 85, 86

I

izlastnoimenski termini 53

K

kalk 54, 55
kalkiranje 54, 55
ključnost 97, 109
konkordanca 30, 31, 42, 91, 94, 96, 97
konverzija 55
korpusna terminologija 19, 43, 84, 90
krajšava 40, 104
kratica 40

L

leksem 19, 20, 24, 33, 34, 42, 59, 94
leksikografi ja 19, 59, 72, 84

leksikologija 9, 19

LISA 74

LL 97, 109

M

MI 85, 107
MultiTerm 63, 76, 114, 115, 116,
117, 119

N

načelo enoznačnosti 18, 27, 37

O

oblika terminov 39
ontologija 28, 31, 33, 34, 35

P

podatkovna kategorija 63
poimenovanje 18, 19, 24, 27, 28, 37,
51, 54
pojavnica 93, 95, 102, 103
pojem 19, 24, 25, 27, 30, 37, 63, 65,
71, 74, 82
pojmovna razmerja 23, 75
pojmovno načelo 25, 27, 32, 68
pojmovno polje 23, 27, 29, 65
pomenotvorje 55
prevzemanje 51, 52, 53, 54, 56, 104

R

različnica 93, 94, 95, 96, 98
razmerja med pojmi 23, 28, 29, 30, 31,
37, 65, 68, 82

S

semantični splet 28, 34
semantično omrežje 25, 31
seznam praznih besed 92, 107
skladenjski vzorec 39
SloWNet 46
strokovni jezik 13, 14, 15, 16, 51

strokovno-specifična referenca 38

T

TBX 62, 63, 74

TEI 73

termin 18, 24, 30, 37, 39, 40, 46, 63,
64, 65, 75, 101, 102, 103, 111, 118

terminografi ja 19, 59, 60, 67, 78, 84

terminolog 47, 99

terminologija 13, 18, 19, 46

terminološka baza 25, 62, 63, 64, 73,
113, 120

terminološka veda 18, 28, 56

terminološke kolokacije 39, 41, 97

terminološke variacije 27, 42, 84, 85

terminološki slovar 41, 48, 59, 61, 67

terminološkost 38, 39

terminotvorje 50, 51, 56

t-test 85

TTR 95

W

Wordsmith 90, 91, 96, 97

X

XML 73, 76, 91, 119

X2 85, 86, 107

Viri



- Ahmad, K., Davies, A., Fulford, H. in Rogers, M., 1992: What is a term? The semi-automatic extraction of terms from text. Snell Hornby, M. (ur.) *Translation Studies*. Amsterdam: John Benjamins.
- Arntz, R. in Picht, H., 1995: *Einführung in die Terminologearbeit*. Hildesheim: Olms.
- Baroni, M., Kilgariff, A., Pomikálek, J., Rychlý, P., 2006: WebBootCaT: instant domain-specific corpora to support human translators. *Proceedings of EAMT 2006*, Oslo. 247-252
- Biber, D., 1993: Representativeness in corpus design. *Literary and Linguistic Computing* 8, 243-257
- Bokal, L., 1998: Tipologija novih besed (ob primeru avtomobilskega izrazja). Humar, M. (ur.) *Slovensko naravoslovno-tehnično izrazje*. Ljubljana: ZRC SAZU, 147-162
- Bourigault, D., Gonzales-Mullier, I. in Gros, C., 1996: Lexter: a natural language processing tool for terminology extraction. *Proceedings of the 7th EURALEX International Congress*, Gothenburg, Švedska.
- Buitelaar, P., Eigner, Th., Declerck, T., 2004: OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection. *Proceedings of the Demo Session at the International Semantic Web Conference*, Hirošima, Japonska.
- Cabré, M. T., 1998: *Terminology: theory, methods and applications*. Amsterdam, John Benjamins.
- Dagan, I. in Church, K., 1994: Termight: Identifying and translating technical terminology. *Proceedings of the Fourth Conference on Applied Natural Language Processing*. 34-40
- Daille, B., 1995: *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Lancaster: UCREL Lancaster University. (URL: <http://www.comp.lancs.ac.uk/ucrel/papers/techpaper/vol5.pdf>)
- Erjavec, T., 1996/97: Računalniške zbirke besedil. *Jezik in slovnstvo* 2/3. 81-95
- Erjavec, T. in Vintar, Š., 2004: Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. *Uporabna informatika* 12/2 (2004). 97-106
- Estopà Bagot, R., 1999: *Extraccio de terminologia: elements per a la construccio d'un SEACUSE (Sistema d'Extraccio Automatica de Candidats a Unitas de Significacio Especialitzada)*. Doktorska disertacija. Barcelona: Univerza Pompeu Fabra.
- Felber, H., 1984: *Terminology Manual*. Paris: Infoterm.
- Fišer, D., 2007: A multilingual approach to building Slovene Wordnet. In: *Proceedings of the workshop on A Common Natural Language Processing Paradigm for Balkan Languages held within the Recent Advances in Natural Language Processing Conference RANLP'07*. Borovets, Bulgaria.
- Galinski, C. in H. Picht, 1995: Graphic and other semiotic forms of knowledge representation in terminology work. Wright, S.E. in Budin, G. *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins.

- Gorjanc, V., 1996: Terminologija novejših naravoslovno-tehniških strok (Ob primeru računalništva in jedrske fizike). Vidovič Muha, A. (ur.): *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 251–260
- Gorjanc, V., 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.
- Gorjanc, V. in Vintar, Š., 2007: Korpusna analiza vloge označevalcev medleksemskih razmerij v organizaciji besedila. *Jezik in slovstvo*, 52/3-4. 117-129
- Gorjanc, V. in Logar N., 2007. Od splošnih do specializiranih korpusov - načela gradnje glede na njihov namen. Orel, I. (ur.) *Razvoj slovenskega strokovnega jezika*. Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, 2007. (Obdobja. Metode in zvrsti ; 24). 637-650
- Heid, U., Evert, S., Fitschen, A., Freese, M., in Vögele, A., 2001: *Term candidate extraction in DOT*. Dot final report, Part II. Stuttgart: IMS, Univerza v Stuttgartu.
- Hiemstra, D., 1998: Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. Coppen, Peter-Arno in dr., (ur.): *Proceedings of the 8th CLIN meeting*. 41-58
- Hočvar, J. in Kiralj, E., 2004: Vodenje terminologije in priprava standardiziranega izrazja na nacionalni ravni. Humar, M. (ur.): *Terminologija v času globalizacije*. Ljubljana: ZRC SAZU. 201-212
- Hoffmann, L., 1985: *Kommunikationsmittel Fachsprache*. Tübingen: Günther Narr Verlag.
- Jacquemin, C., 2001: *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Massachusetts: MIT Press.
- Justeson, J.S. in Katz, S., 1995: Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1/95. 9-27
- Kageura, K., 2002: *The Dynamics of Terminology. A descriptive theory of term formation and terminological growth*. Amsterdam/Philadelphia: John Benjamins.
- Kalin Golob, M., 2001: Tvorjenje komunikološkega izrazja ob prevajanju temeljnih komunikoloških del. Splichal, S. (ur.) *Komunikološka hrestomatija 1. Začetki komunikologije v Evropi in ZDA*. Ljubljana: Fakulteta za družbene vede. 251-260
- Kilgariff, A., 2001: Comparing Corpora. *International Journal of Corpus Linguistics* 6 (1). 1-37
- Kilgariff, A., Rychly, P., Smrz, P. in Tugwell, D., 2004: The Sketch Engine, *Proceedings Euralex*, Lorient, Francija. 105-116
- Klinar, S., 2004: Tehnični termin, kaj je to? Humar, M. (ur.): *Terminologija v času globalizacije*. Ljubljana: ZRC SAZU. 103-109
- Korošec, T., 1998: *Slovenski vojaški jezik*. Ljubljana: Fakulteta za družbene vede.
- Košmrlj-Levačič, B., 1998: Izimensko strokovno izrazje. Humar, M. (ur) *Slovensko naravoslovno-tehnično izrazje*. Ljubljana: ZRC SAZU. 139-146

- Košmrlj-Levačič, B. in Seliškar, T., 2004: Uporabniški računalniški program SlovarRed 2.0. Humar, M. (ur.): *Terminologija v času globalizacije*. Ljubljana: ZRC SAZU. 179-199
- Logar, N., 2005: Norma v slovarju sodobne slovenščine: zloženke in kratice. *Družboslovne razprave*, XXI (2005), 48. 211-225
- Losee, R. M., 1995: The development and migration of concepts from donor to borrower disciplines: Sublanguage term use in hard & soft sciences. *Proceedings of the 5th International Conference on Sociometrics and Infometrics*, Chicago, ZDA. 265-274
- Malaisé, V., Zweigenbaum, P. in Bachimont, B., 2007: Mining defining contexts to help structuring differential ontologies. Ibekwe-SanJuan in dr. (ur.) *Application-Driven Terminology Engineering*. Amsterdam, Philadelphia: John Benjamins. 19-48.
- Manning, C. in Schütze, H., 2000: *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.
- Melby, A. K., 2003: *Interchange using TBX*. LISA Oscar Meeting, Washington DC, Brigham Young University, Provo Campus (URL: www.lisa.org/sigs/terminology/tbx_intro/)
- Och, F. J., Ney, H., 2003: A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29/1. 19-51
- Pearson, J., 1998: *Terms in Context*. Amsterdam, Philadelphia: John Benjamins.
- Pisanski, A., 2002: Analiza nekaterih metabesedilnih elementov v slovenskih znanstvenih člankih v dveh časovnih obdobjih. *Slavistična revija* 50/2. 183-197
- Puc, K., Erjavec, T., 2006: Uporaba korpusa pri urejanju spletnega terminološkega slovarja. Erjavec, T. in Žganec Gros, J. (ur.) *Language technologies/Jezikovne tehnologije IS-LTC*. Ljubljana: Institut Jožef Stefan. 156-161
- Sager, J. C., 1998/99: In search of a foundation: Towards the theory of the term. *Terminology* 5(1). 41-57
- Toporišič, J., 1984: *Slovenska slovnica*. Maribor: Obzorja.
- Toporišič, J., 2006: O našem zdravstvenem jeziku. Prvič objavljeno v: Medicinski razgledi 7 (1968: 429-442). *Besedjeslovne razprave*. Ljubljana: ZRC SAZU. 449-461
- Vidovič Muha, A., 2000: *Slovensko leksikalno pomenoslovje: Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.
- Vintar, Š., 2002: Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. *Zbornik konference ISJT'02*. Ljubljana: Institut Jožef Stefan.
- Vintar, Š., 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija. Univerza v Ljubljani: Filozofska fakulteta.
- Wright, S. E. in Budin, G., ur., 1995: *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins.

- Whorf, B. (Carroll, J., ur.), 1956: *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge: MIT Press.
- Zielinski, D. in Ramírez, Y., 2005: *Research meets practice: t-survey 2005*. An online survey on terminology extraction and terminology management. (URL: fr46.uni-saarland.de/download/publs/sdv/t-survey_aslib2005_zielinski.htm)
- Žele, A., 2004: Stopnje terminologizacije v leksiki (na primerih glagolov). Humar, M. (ur.): *Terminologija v času globalizacije*. Ljubljana: ZRC SAZU. 77-91

THE RIGHT