

Vairākvārdu leksēmu klasificēšana elektroniskajā vārdnīcā „Tēzaurs”¹

Classification of multi-word units in the electronic dictionary „Tēzaurs”

Gunta Nešpore-Bērzkalne, Ilze Lokmane

Matemātikas un informātikas institūts

Latvijas Universitāte

Raiņa bulvāris 29, Rīga, LV-1050, Latvija

E-pasts: gunta.nespore@lumii.lv, ilze.lokmane@lu.lv

„Tēzaurs” ir lielākā latviešu valodas elektroniskā vārdnīca, kurā ir vairāk nekā 400 000 šķirklju, tostarp vairāk nekā 71 000 ir vairākvārdu leksēmu (VVL) šķirklju, kas nākuši no visdažādākajiem avotiem un ir ļoti daudzveidīgi gan pēc satura, gan formas. Pēdējos gados notiek darbs pie „Tēzaurā” atrodamo VVL klasificēšanas vairākās grupās: frazēmas, idiomas, kolokācijas, vārdkoptermini, taksoni, vairākvārdu vietvārdi. Šajā rakstā aprakstīti VVL līdzšinējās klasificēšanas rezultāti, šī darba laikā radušās grūtības, kas pamatā saistītas ar izvēlēto VVL grupu robežošanas un pārklāšanas, kā arī ar nepietiekami precīzi definētiem grupu nošķiršanas kritērijiem. Konstatēts, ka vajadzīgi papildu kritēriji frazēmu un idiomu, kā arī frazēmu un kolokāciju nošķiršanai.

Ieguvumi no VVL iedalījuma mazākās grupās ir dažādi. Vārdnīcās lietotājiem šis papildu marķējums palīdz labāk orientēties vārdnīcas materiālā. „Tēzaura” veidotāji ir ieguvuši plašāku informāciju par VVL kopuma sastāvu, iespējas analizēt viena tipa vārdu savienojumus vienkopus, novērst informācijas dublēšanos un neprecizitātes, uzlabojot vārdnīcas kvalitāti, iespēju pārskatīt un precizēt VVL novietojumu vārdnīcas šķirklī. Valodniekiem apkopotais un grupētais valodas materiāls paver iespējas veikt katras VVL grupas dziļāku izpēti un norāda nepieciešamos pētījumu virzienus.

Pašreizējā VVL klasifikācija ir pirmais solis sakārtotas VVL apraksta sistēmas virzienā, kas nākotnē ir jāpārskata, jāprecizē un jāpapildina.

Atslēgvārdi: vairākvārdu leksēma; frazeoloģija; idioma, frazēma; kolokācija; vārdkoptermins, elektroniskā vārdnīca „Tēzaurs”.

¹ Pētījums sagatavots Latvijas Zinātnes padomes fundamentālo un lietišķo pētījumu projektā „Latviešu valodas skaitļojamo leksisko resursu attīstīšana dabiskās valodas sapratnei un tekstradei” (Nr. lzp-2022/1-0443). Izsakām pateicību kolēģēm Laurai Ritumai, Agutei Klintij un Madarai Stādei par ieguldījumu raksta tapšanā.

levads

„Tēzaurs” ir lielākā latviešu valodas elektroniskā vārdnīca (Spektors *et al.* 2016), kas veidota kopš 20. gs. 90. gadu sākuma, digitalizējot un apvienojot visas pieejamās latviešu valodas vārdnīcas un citus leksikas krājumus (vairāk nekā 340 avotu), tās papildināšana un datu kārtošana turpinās. Vārdnīcā ir aptuveni 400 000 šķirklī, no kuriem vairāk nekā 71 000 ir vairākvārdu leksēmu (turpmāk – arī VVL) šķirklī.

„Tēzaura” vajadzībām tiek izmantota Lorijs Bauera (*Laurie Bauer* 2022, 5) piedāvātā vairākvārdu leksēmas (ang. *multi-word expression* vai *multi-word unit*) izpratne, proti, vairākvārdu leksēma ir vārdnīcas lemma, ko veido vairāk nekā viens ortogrāfisks vārds. Kā ekvivalents minētajiem angļu terminiem izvēlēts apzīmējums *vairākvārdu leksēma*, jo no pārējām „Tēzaura” lemmām (resp., leksēmām, kas sastāv no viena vārda) vairākvārdu leksēmas atšķiras pamatā formāli – pēc vārdu skaita. Tā ir jebkura vairākvārdu lemma, kurai vārdnīcā ir definīcija. VVL var būt ne vien vairāku pilnnozīmes vārdu savienojumi (1), bet arī vietniekvārdi (2), apstākļa vārdi (3), vairākvārdu saikļi (4) un partikulas (5).

- (1) *slinkuma maiss, labot kļūdas, baltum balts*
- (2) *diez kas, kaut kāds*
- (3) *nezin kur, tik tikko*
- (4) *kā arī, tā kā*
- (5) *diez vai, vai nu ne*

Arī saturiski VVL ir ļoti dažādas – ģeogrāfiskie nosaukumi, termini, frazeoloģismi, dažādi izteicieni – jebkādi vārdu savienojumi, ko kādas vārdnīcas veidotāji uzskatījuši par skaidrošanas vērtiem un iekļāvuši vārdnīcā. Tāpat kā pārējām leksēmām VVL ir savs šķirklis ar lemmu, nozīmes skaidrojumu, lietojuma piemēriem, ja tādi pievienoti, un citu informāciju.

Vairākvārdu leksēma parasti ir piesaistīta viena vai vairāku tās sastāvā esošo vārdu šķirklīem (piem., *slinkuma maiss* ir norādīts gan šķirklī *slinkums*, gan šķirklī *maiss*), bet viens vārds var būt iesaistīts vairākās VVL, tāpēc vienā vārdnīcas šķirklī var būt minēti vairāki desmiti un pat simti VVL (piem., šķirklī *acs* ir 207 vairākvārdu leksēmas). „Tēzaura” šķirklī tās atrodamas sadaļā „Stabili vārdu savienojumi”. Sākotnēji „Tēzaurā” VVL nebija sīkāk klasificētas, tomēr nepieciešamība sistematizēt milzīgo datu apjomu bija neapšaubāma – tas palīdz apzināt „Tēzaurā” iekļauto VVL daudzveidību un dažādu VVL grupu apjomu, kā arī sniegt papildu informāciju par šīm leksēmām.

Dažādu valodu elektroniskajos resursos VVL dalījums grupās ir atšķirīgs, jo atšķiras gan pašas valodas, gan valodu apraksta tradīcijas. Piem., pēc angļu valodas elektronisko resursu parauga arī citās valodās vairākvārdu leksēmu kopā kā atsevišķas grupas tiek iekļautas konstrukcijas ar vispārinātas nozīmes desemantizētiem darbības vārdiem (piem., ang. *light verb constructions*, kā *take part* ‘piedalīties’, *make a mistake* ‘kļūdīties; pieļaut kļūdu’) un darbības vārdu un postpozīciju savienojumi (ang. *adpositional verbs*, piem., *give up* ‘padoties’, *care about* ‘rūpēties’) (Leseva *et al.* 2020). Latviešu valodā ir līdzīgas parādības (6), tomēr tās tradicionāli netiek uzskatītas par īpašām vairākvārdu leksēmu grupām, tāpēc arī šajā rakstā aprakstītajā klasifikācijas

sistēmā tās netiek nošķirtas atsevišķi, bet ietilpst kolokāciju grupā (tuvāk par šo grupu sk. 1. nodaļā).

(6) *darīt zināmu, izdarīt ierakstu, ķerties klāt, ņemt nost*

Šī raksta mērķis ir aprakstīt VVL klasificēšanu „Tēzaura” materiālā, izvēlēto VVL grupu nošķiršanas problēmas, kas radušās darba laikā, un iespējamās to risinājumus, kā arī līdzšinējos VVL klasificēšanas rezultātus.

1. Vairākvārdu leksēmu klasificēšana „Tēzaurā”

Pēdējos gados vairākos piegājienos ir notikusi „Tēzaura” VVL klasificēšana – 2022. gadā tika nošķirti vairākvārdu vietvārdi un starptautiskie (latīņu cilmes) taksoni (sugu, dzimšu, kārtu u. c.) nosaukumi: attiecīgs marķējums (atzīme pie VVL leksēmas, ka tā pieder pie noteiktas grupas) tika pievienots 14733 vietvārdiem un 10347 starptautiskajiem taksonu nosaukumiem (Grasmanis *et al.* 2023). Šajā rakstā tuvāk raksturots „Tēzaura” VVL klasificēšanas nākamais posms, kas notika 2023. gada rudenī un kura mērķis bija klasificēt pārējās VVL, īpašu uzmanību pievēršot frazeoloģismu nošķiršanai no citām vairākvārdu leksēmām un mēģinot tās sīkāk iedalīt idiomās un frazēmās. Šajā darbā piedalījās 12 Latvijas Universitātes Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijas (AiLab) pētnieki. Apstrādājamā materiālā tika iekļautas apmēram 46 000 leksēmas: visas „Tēzaurā” esošās VVL, izņemot, pirmkārt, vietvārdus un starptautiskos taksonu nosaukumus, kas jau bija marķēti iepriekš, otrkārt, citus īpašvārdus (VVL, kas sākas ar lielo burtu vai kas ir marķētas kā īpašvārdi), treškārt, vārdu savienojumus svešvalodā (piem., latīņu spārnotos teicienus vai mūzikas terminus itāliešu valodā), kuri arī jau ir attiecīgi marķēti.

Vispārīnnot vārdnīcas empīrisko materiālu, „Tēzaura” veidotāji izveidoja šādu VVL klasifikāciju ar atbilstošām grupu definīcijām:

- *Vārdkoptermins* – vārdu savienojums, kas apzīmē jēdzienu kādā nozarē un kam ir specializēta nozīme un lietošanas joma (Skujiņa 2007, 399–400, 435), un kas ir oficiāli apstiprināts, atrodams terminu datubāzē vai tamlīdzīgā avotā, piem.:
- (7) *skaņu raksts, iekšzemes kopprodukts*
- Darba gaitā šī grupa tika paplašināta, iekļaujot tajā arī terminu kandidātus – vārdu savienojumus ar termina pazīmēm, kas nav atrodami terminu datubāzē, t. i., var nebūt oficiāli apstiprināti.
- *Vairākvārdu vietvārds* – īpašvārds, kas nosauc ģeogrāfisku objektu (Skujiņa 2007) un kas sastāv no vismaz diviem vārdiem, piem.:
- (8) *Babītes ezers, Vidzemes augstiene*
- *Taksons* – sugas, ģints, dzimtas, kārtas, klases vai tipa nosaukums (Baldunčiks 1999, 775) latviešu valodā, piem.:
- (9) *blāvā akācija, purva piekūns*

Sākotnēji šī grupa bija plānota kā tematiska terminu apakšgrupa, bet darba gaitā kļuva plašāka – tajā ir iekļauti visi ar vārdkopām izteiktie augu un dzīvnieku nosaukumi un to varianti, arī izlokšņu un sarunvalodas nosaukumi, kas nav uzskatāmi par terminiem, piem.:

(10) *dzīvības koks, dzeguzes piestiņi*

- *Frazeoloģisms* – relatīvi stabils ar valodas tradīciju nostiprināts vārdu savienojums, kura nozīme saistīta ar visa vārdu savienojuma vai atsevišķu tā komponentu nozīmes pārnese (Skujiņa 2007, 124). Frazeoloģismi iedalāmi sīkāk:
 - *idioma* – frazeoloģisms, kurā neviens komponents nav lietots savā leksiskajā nozīmē (Skujiņa 2007, 148), piem.:

(11) *jūra līdz ceļiem; kaķis maisā*

- *frazēma* – frazeoloģisms, kurā viens komponents lietots savā leksiskajā nozīmē (Skujiņa 2007, 123), piem.:

(12) *domu grauds, azbesta vecene*

- *Kolokācija* (ang. *collocation*) – bieži lietots vārdu savienojums bez nozīmes pārnese (Matthews 1997, 60), piem.:

(13) *izdarīt ierakstu*

Vairākvārdu leksēmu klasificēšanas vajadzībām LUMII Mākslīgā intelekta laboratorijā tika izstrādāts rīks, ar kuru tās var ērti iekļaut kādā no grupām: frazeoloģismos (ja iespējams, tiek norādīts frazeoloģisma tips – frazēma vai idioma), kolokācijās, taksonos vai vārdkopterminos.

Rīkā (sk. 1. attēlu) kreisajā pusē redzams nemarkēto vairākvārdu leksēmu saraksts (vienā lapā līdz 1000 VVL), katrai klasificējamai vairākvārdu leksēmai apakšā redzama tās definīcija, kā arī papildinformācija par leksēmu, ja vārdnīcā tāda ir, piem., norāde par jomu, kādā leksēma tiek lietota, vai lietojuma ierobežojumiem. Labajā pusē katrai leksēmai pretī redzamas pogas, uz kurām ir VVL klasifikācijā izmantoto grupu apzīmējumi. Noklikšķinot uz VVL grupas apzīmējuma, vārdnīcas datubāzē leksēmai tiek pievienots attiecīgs marķējums, piem., ja noklikšķina uz pogas ar apzīmējumu „F+Id”, vairākvārdu leksēmai tiek pievienota informācija, ka tā ir frazeoloģisms un ka šī frazeoloģisma tips ir idioma. Marķētās vairākvārdu leksēmas no saraksta pazūd, lai vieglāk pārskatīt vēl nemarkētās VVL.

lokalizētā amnēzija • atmiņas zudums par noteiktu laika posmu		F+Id F+Fr F Kol Tks Vkt
lokāmais divdabis • divdabis, kam ir darbības vārda kārtas, laika un veida kategorija un īpašības vārda dzimtes, skaitļa, locījuma, daļēji - arī noteiktības un gradācijas kategorija; deklinējama particips		{"Joma":["Valodniecība"]} F+Id F+Fr F Kol Tks Vkt
lokās kā teļa aste • cenšas izlocīties, izmēloties, izkļūt no kādas nepatīkamas situācijas		{"Lietojums":["Apvīdārdš"]} F+Id F+Fr F Kol Tks Vkt

1. attēls. Vairākvārdu leksēmu klasificēšanas rīks

2. Vairākvārdu leksēmu klasificēšanas rezultāti

Līdz 2023. gada beigām tika klasificēta lielākā daļa no „Tēzaura” vairākvārdu leksēmām – no 71304 VVL marķētas ir 65919 VVL (šeit un citur norādīti 2023. gada decembra dati; tā kā vārdnīca tiek pastāvīgi rediģēta un papildināta, laika gaitā šie skaitļi var nedaudz mainīties).

Klasificēto „Tēzaura” vairākvārdu leksēmu skaits un grupu proporcijas redzamas 1. tabulā:

VVL grupa	VVL apakšgrupa	VVL skaits	VVL īpatsvars
Vārdkoptermini		22552	31,6 %
Vairākvārdu vietvārdi		14733	20,7 %
Taksoni	Starptautiskie nosaukumi	10347	14,5 %
	Latviešu nosaukumi	7854	11,0 %
Frazeoloģismi	Frazēmas	2863	4,0 %
	Idiomas	4029	5,7 %
	Tips nav norādīts	358	0,5 %
Kolokācijas		3183	4,5 %
Klasificētās VVL kopā		65919	92,5 %
Neapstrādātās VVL		5385	7,6 %
VVL kopskaits „Tēzaurā”		71304	100 %

1. tabula. Vairākvārdu leksēmu skaits katrā grupā un grupas īpatsvars

Vislielāko VVL grupu veido termini (31,6 %) – tas skaidrojams ar to, ka „Tēzaurā” iekļauts ne tikai vispārīgo skaidrojošo vārdnīcu materiāls, bet arī termini no visām latviešu valodā izdotajām skaidrojošajām terminoloģijas vārdnīcām („Tēzaura” avotu sarakstu sk. https://tezaur.lv/_avoti). Līdzīgu iemeslu dēļ lielas grupas ir arī vairākvārdu vietvārdi (20,7 %), kas marķēti iepriekš, un taksoni (starptautiskie nosaukumi 14,5 %, latviešu nosaukumi 11,0 %, kopā 25,5 %), jo „Tēzaurā” iekļauti dati no dažādām ģeogrāfisko nosaukumu vārdnīcām, vietvārdu krājumiem, augu nosaukumu krājumiem u. tml. Vismazākā grupa savukārt ir kolokācijas (4,5 %), iespējams, tāpēc, ka vārdnīcās retāk iekļauj vārdu savienojumus, kam nav ne nozīmes pārnese, ne arī kādu specifisku (terminoloģisku) funkciju. Kā redzams tabulā, ir iegūta arī apjomīga (frazēmas, idiomas un sīkāk nediferencēti frazeoloģismi – kopā 7250 vienību), lai arī proporcionāli ne tik liela (kopā 10,2 %) frazeoloģismu grupa, kas paver iespējas tos pētīt detalizēti, ko būtu grūti darīt, ja frazeoloģismi nebūtu nošķirti no pārējām VVL.

3. Vairākvārdu leksēmu klasifikācijā neiekļautās grupas

Marķējot „Tēzaura” datus, tika konstatētas vairākas VVL grupas, kuras neiederas izveidotajā klasifikācijā.

Neliela daļa izteicienu svešvalodā „Tēzaura” datos vēl nebija attiecīgi marķēti (14; 15) – tiem pievienots svešvalodas marķējums, bet VVL svešvalodā šobrīd paliek ārpus definētajām grupām.

- (14) *laterna magica* ‘aparāts attēlu projicēšanai, izgudrots, domājams, 17. gs.’
 (15) *opera buffa* ‘komiskā opera’

Nākotnē var apsvērt iespēju izveidoto klasifikāciju attiecināt arī uz VVL svešvalodā. Šajā pašā grupā iekļaujami arī latviski pierakstīti neliterāri citvalodu izteicieni, piem.:

- (16) *baba jaga* ‘ragana; slota’
 (17) *jo majo* ‘izsauciens, kas izsaka pārsteigumu’
 (18) *gud bai* ‘saka, kā atvadu sveicienu’

Ārpus klasifikācijas šobrīd paliek arī saīsinājumi, kas sastāv no vairākiem elementiem, piem.:

- (19) *op. cit., a. god., t. sk.*

Šādām VVL jāveido pilna vārdu savienojuma šķirklis, pie kura pievienojama saīsinājuma leksēma (piem., ir jāizveido šķirklis *augsti godātais*, kurā iekļauta leksēma *a. god.*), tad saīsinājumi tiks klasificēti kopā ar nesaīsināto vārdu savienojumu. Daļa šādu šķirklju jau ir izveidota, darbs pie to veidošanas turpinās.

Atsevišķu grupu veido šķirklī, kuru lemma ir nosaukums ar nomenklatūrvārdu, piem.:

- (20) *ātrvilciens Eurostar; datne README, operētājsistēma NetWare*

Šie šķirklī ir pārstrādājami tā, lai lemma būtu nosaukums bez nomenklatūrvārda, bet nomenklatūrvārds parādītos tikai nozīmes skaidrojumā. Tādā gadījumā šīs vairs nebūs VVL, bet pagaidām arī šie šķirklī paliek ārpus klasifikācijas.

Nevienā no pašreizējās klasifikācijas grupām neiederas dažādas nosaucējvienības – mitoloģijas tēlu, rotaļu, deju, mēnešu u. tml. nosaukumi –, kas tradicionāli netiek iekļauti frazeoloģismos, piem.:

- (21) *dieva dēli, babiloniešu netikle, čūsku ķēniņš, sietiņa dzišana, lapu mēnesis, garais Ancis*

Neklasificēta pagaidām atstāta vēl viena daudzveidīga nosaucējvienību grupa – ēdienu nosaukumi, piem.:

- (22) *viltotais zaķis, jauktie kami, kafijas krējums*

Šajā grupā sastopami visdažādākie tālāk aprakstīto klasificēšanas problēmu piemēri jeb gadījumi, kad vārdu savienojums iekļaujas vairākās grupās (piem., tas ir vienlaikus termins un idioma). Turklāt kulinārijas un pārtikas tehnoloģijas termini ir samērā reti sastopami terminu datubāzēs vai vārdnīcās (Škapare 2020), tāpēc trūkst informācijas terminu nošķiršanai no neterminoloģiskiem vārdu savienojumiem. Nākotnē šo tematisko grupu varētu izmantot papildinātas klasifikācijas izstrādei un testēšanai.

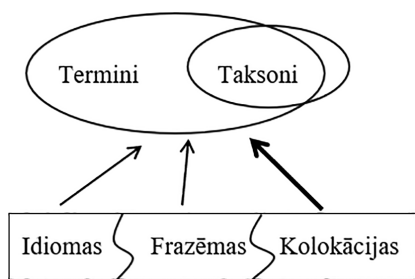
4. Vairākvārdu leksēmu klasificēšanas problēmas un risinājumi

Tā kā izmantotā dalījuma pamatā ir dažādi kritēriji – pirmkārt, pēc tēlainības (frazeloģismi) vai tās trūkuma (kolokācijas), otrkārt, pēc funkcijām un lietojuma jomas (termini), treškārt, pēc tematikas (taksoni), grupu pārklāšanās un robežošanās ir neizbēgama.

Sākotnējā klasifikācija pēc jauktiem kritērijiem tika izveidota praktisku apsvērumu dēļ – lai klasificēšanu īsā laikā varētu organizēt kā talku, kuras dalībniekiem ir jāizvēlas tikai viena iespējami skaidri definēta grupa, kurā vairākvārdu leksēmu iekļaut. Mērķis bija īsā periodā apstrādāt pēc iespējas vairāk VVL, lai gūtu labāku priekšstatu par „Tēzaurā” esošajām VVL un pēc tam varētu labāk plānot tālāko VVL kārtošanas un apstrādāšanas procesu. Tā kā „Tēzaurā” ir iestrādātas visas publicētās skaidrojošās terminu vārdnīcas un terminu tajā ir ļoti daudz, tie tika nošķirti atsevišķā grupā, pieņemot, ka termini būs salīdzinoši viegli nosakāmi, meklējot informāciju „Latvijas Nacionālajā terminoloģijas portālā” (turpmāk – LNTP; pieejams: <https://termini.gov.lv/>). Atsevišķi tika izveidota tematiski viegli nošķiramā taksonu grupa, kaut gan daļa taksonu ir terminu apakšgrupa. Tā kā frazeoloģismu dalījums divās grupās atkarībā no tā, vai nozīmes pārnese ir pilnīga vai daļēja, latviešu valodniecībā ir tradicionāls (Laua 1992, 24; Skujiņa 2007, 123, 148; Veisbergs 2013b, 157), „Tēzaura” veidotāji nolēma paturēt dalījumu frazēmās un idiomās, pirmkārt, lai daudzveidīgā frazeoloģismu grupa būtu strukturēta, un, otrkārt, lai šī dalījuma lietderību un pamatotību pārbaudītu praksē. Kolokāciju grupā savukārt bija paredzēts iekļaut vārdu savienojumus, kas pārstāj grupās neiederas, lai vēlāk šo grupu analizētu un, iespējams, dalītu sīkāk.

Tika konstatēta gan VVL grupu pārklāšanās, gan robežošanās, kas shematiski parādīta 2. attēlā.

Par pārklāšanos jārunā gadījumos, kad, vērtējot pēc dažādiem kritērijiem, vairākvārdu leksēma vienlaikus iekļaujas vairāk nekā vienā grupā (piem., gan termins, gan taksons var būt tēlaini, tātad atbilst frazeoloģisma pazīmēm; liela daļa taksonu ir termini). Lai VVL iedalījums būtu pēc iespējas konsekventāks, darba gaitā tika definēts, kurai grupai šādos gadījumos dodama priekšroka: 1) taksons (arī tad, ja tas vienlaikus ir termins, kolokācija vai frazeoloģisms), 2) termins (arī tad, ja tas vienlaikus ir kolokācija vai frazeoloģisms), 3) idioma, frazēma, kolokācija (tikai tad, ja tie nav ne taksoni, ne termini).



2. attēls. Attieksmes starp VVL grupām

Šāda marķēšanas secība izvēlēta, pirmkārt, lai vispirms tiktu marķētas skaidrākās grupas – taksoni pēc denotāta parasti ir viegli nosakāmi, arī termini vairumā gadījumu ir konstatējami pēc fiksējuma attiecīgās vārdnīcās vai datubāzēs –, otrkārt, „Tēzaura” veidotāji uzskata, ka vārdnīcas lietotājiem lietderīgāk būtu zināt, ka attiecīgais vārds ir kādas nozares termins neatkarīgi no tā, vai tajā ir nozīmes pārnese vai nav. Savukārt robežošanās vērojama starp grupām, kurās leksēmas ir iedalītas pēc tēlainības vai tās trūkuma, – tādas ir idiomas, frazēmas un kolokācijas.

VVL klasificēšanas laikā tika konstatētas vairākas problēmas un nepietiekami skaidri definēti grupu nošķiršanas kritēriji, kuru dēļ ne vienmēr bija skaidrs, kurā grupā attiecīgā leksēma iekļaujama. Vismazāk grūtību sagādāja taksonu marķēšana, bet neskaidrības radās galvenokārt šādu grupu robežu noteikšanā: 1) vai VVL ir termins vai kolokācija, 2) vai VVL ir frazeoloģisms vai kolokācija, 3) vai VVL ir frazēma vai idioma.

Šīs marķēšanas grūtības ir aprakstītas turpmāk.

4.1. Terminu marķēšana

Termins tiek definēts kā „konkrētas tematiskās jomas, speciālās nozares jēdziena vārds disks apzīmējums — vārds vai vārdkopa, kas izteic (nosauc un iezīmē) attiecīgās jomas (nozares) noteiktu jēdzienu kā nozares jēdzienu sistēmas sastāvdaļu” (Skujiņa 2007, 399–400). Tomēr noteiktā profesionālās darbības jomā lietoti vārdu savienojumi, kas nav iekļauti LNT, nav droši nošķirami no profesionāļiem (šajā klasifikācijas sistēmā – kolokācijām), jo tie nav oficiāli pieņemti termina statusā, tomēr ir viens no zinātniskās terminoloģijas papildināšanas avotiem (Skujiņa 2007, 312–313).

Lai gan sākotnēji bija paredzēts terminu grupā iekļaut tikai oficiāli apstiprinātos un LNT iekļautos terminus, talkas gaitā tika konstatēts daudz vārdu savienojumu, kas gan pēc lietojuma norādēm, gan jēdziena skaidrojuma līdzinās terminiem, tomēr LNT nav atrodam. Spilgts piemērs ir jurniecības jēdzieni, kas „Tēzaurā” ir iekļauti no „Jurniecības vārdnīcā” (Millers 2007). Šajā vārdnīcā ir definēti dažādi ar jurniecību un kuģiem saistīti jēdzieni, kuru oficiālais statuss (termins vai profesionālisms) nav norādīts, savukārt LNT ir iekļauta tikai neliela daļa no tiem. Piem., vairākvārdu leksēmas *buras gredzens*, *buru trimmēšana*, *buru kuģis* un *buru laiva* ir iekļautas „Jurniecības vārdnīcā”, bet LNT no šīm ir atrodama tikai *buru laiva*.

Līdzīgi arī ar mitoloģiju saistītu jēdzienu un ēdienu nosaukumi LNT sastopami reti. Tas nenozīmē, ka attiecīgajās jomās terminoloģija nepastāv, bet tā, visticamāk, nav sistematizēta un oficiāli apstiprināta. Šādas leksēmas uzskatāmas par terminu kandidātiem, tāpēc daudzas no tām tika iekļautas terminu grupā. Faktiski terminu kandidātu un kolokāciju robeža ir grūti nosakāma, un leksēmām piešķirtais marķējums ne vienmēr ir konsekvents. Kā minēts iepriekš, ārpus klasifikācijas šobrīd ir atstāti ēdienu nosaukumi, mitoloģijas tēli un līdzīgas nosaucējvienības, jo bez papildu izpētes ir grūti noteikt, kuras no tām ir termini vai terminu kandidāti, bet kuras – neterminoloģiski vārdu savienojumi, tātad kolokācijas, piem., vai *kartupeļu biežputra* un *viltotais zaķis* ir kulinārijas / pārtikas tehnoloģijas termini vai nav.

Termini var būt arī tēlaini, piem.:

- (23) *auna piere* 'ledāju nogludināts cietu iežu izcilnis vai paugurs'
- (24) *auss gliemene* 'garš, spirālveidīgs, ar šķidrumu pildīts kanāls dzirdes orgāna iekšējā daļā'

Tas pats attiecas uz taksoniem, piem.:

- (25) *apšu kundziņš* 'apšu beka'
- (26) *kumeļa pauti* 'parastā bisīte'

Kaut gan Alīse Laua (1992, 27) uzsver, ka nedz salikti termini, nedz salikti nosaukumi nav uzskatāmi par frazeoloģismiem, pat ja tie ir tēlaini, tomēr šāda lēmuma motivācija nav īsti skaidra un būtu apstrīdama. Arī Andrejs Veisbergs (2013a, 115) uzskata, ka tāpēc vien, ka vārdu savienojums ir ieguvis termina statusu, tas nezaudē savu frazeoloģisma dabu. Tiesa, arī VVL klasifikācijā funkcija uzskatīta par galveno kritēriju un metaforiski vārdkoptermini iekļauti terminu grupā, tomēr nākotnē klasifikācijas sistēma būtu jāpapildina tā, lai varētu piešķirt dubultu marķējumu, proti, terminiem un taksoniem būtu pievienojamas arī norādes par nozīmes pārnēsumu.

4.2. Frazeoloģismu un kolokāciju marķēšana

Frazeoloģismus raksturo trīs galvenās pazīmes: 1) vairākvārdu struktūra, 2) tradicionālums, 3) pārnesta nozīme, kas raksturīga vārdu savienojumam kopumā un neizriet no atsevišķo komponentu nozīmēm (Laua 1992, 20; Veisbergs 2019, 114). Pirmo pazīmi iespējams papildināt ar norādi, ka vārdi, kas veido frazeoloģismu, ir sintaktiski saistīti regulārā veidā, tātad no sintaktiskā saistījuma un gramatisko formu viedokļa frazeoloģismā nav nekā neparasta, neprognozējama ir vienīgi nozīme (Mel'čuk 2012, 32). Pirmās divas pazīmes ir vienlīdz attiecināmas arī uz kolokācijām, tāpēc būtiskā frazeoloģismu atšķirība ir nozīmes pārnēsums. Tomēr ne frazeoloģismi, ne kolokācijas nozīmes pārnēsuma ziņā nav viendabīgas, no kā izriet, pirmkārt, frazeoloģismu tipu (idiomu un frazēmu) robežošanās, otrkārt, frazeoloģismu un kolokāciju robežošanās.

Idiomām raksturīgs pilnīgs pārsemantizējums, kas izriet no vārdu savienojuma kā veseluma (27, 28), turpretim frazēmā viens komponents saglabā savu leksisko nozīmi (29, 30) (Laua 1992, 24) un parasti tiek izmantots arī frazēmas nozīmes skaidrojumā.

- (27) *ar gariem zobiem* 'negribīgi'
- (28) *pa galvu, pa kaklu* 'steigā'
- (29) *zili brīnumi* 'lieli brīnumi'
- (30) *slinks kā maiss* 'ļoti slinks'

Turklāt svarīgi, ka pārējo frazēmas komponentu specifiskā nozīme realizējas tikai attiecīgajā vārdu savienojumā, proti, *zils* un *kā maiss* ārpus minētajiem frazeoloģismiem nenozīmē 'liels' vai 'ļoti'.

Šāda frazēmas interpretācija nav pietiekami precīza, lai praktiski nošķirtu frazēmas no idiomām, jo savā leksiskajā nozīmē lietotais vārds var būt: 1) pamatnozīmē (kā A. Lauas minētajos piemēros), 2) atvasinātā tiešā nozīmē, piem., (31), kur vārdam

laime ir atvasināta nozīme ‘veiksme’ (Spektors *et al.*, 2023) 3) atvasinātā pārnestā nozīmē, piem., (32), kur vārdam *nervs* varētu saskatīt pārnestu nozīmi ‘psihiskā izturība, pacietība’ (Spektors *et al.*, 2023), savukārt vārdam *dzelzs* atvasinātu tiešu nozīmi ‘spēcīgs, ļoti izturīgs’ (Spektors *et al.*, 2023).

(31) *cūkas laime* ‘necerēta, negaidīta veiksmē’

(32) *dzelzs nervi* ‘saka par spēcīgu, stipru, stingri apvaldītu psihisko darbību’

Abos pēdējos gadījumos frazeoloģismi drīzāk būtu iekļaujami idiomu grupā, jo tēlainība izriet no vārdu savienojuma kopumā, nevis no komponentu nozīmēm. Frazēmu un idiomu šķīrums praksē izrādījās nepietiekami konsekvents, jo marķētāju izjūta par to, vai vārds ir saglabājis savu leksisko nozīmi, atšķirās atkarībā no tā, cik ļoti šī nozīme ir attālinājusies no pamatnozīmes. Šī iemesla dēļ marķēšanas gaitā iespēju robežās tika ņemts vērā, vai nozīme, kurā attiecīgais vairākvārdu leksēmas vārds ir lietots, ir fiksēta „Tēzaurā”, tomēr šāds kritērijs nešķiet pietiekams konsekventa šķīruma izveidei, jo tā frazēmu grupā nonāk VVL ar ļoti atšķirīgu nozīmes pārnese pakāpi. Raksta autore uzskata, ka nākotnē, ja šāds frazeoloģismu dalījums tipos tiek paturēts, frazēmas definīcija ir precizējama. Viens iespējamais risinājums ir frazēmas izpratni sašaurināt līdz tam, ka viens komponents lietots „pamatnozīmē”, nevis „savā leksiskajā nozīmē”, savukārt otra iespēja ir izvirzīt papildu kritērijus frazeoloģismu tipu nošķiršanā.

Kolokācijas ne vienmēr iespējams stingri nošķirt no frazeoloģismiem, jo tās arī citās valodās mēdz būt vismaz daļēji idiomātiskas (Burger 2007, 101–102). Līdzās neapšaubāmām kolokācijām, kur visi vārdi lietoti pamatnozīmē (33), ir tādas, kurās ietilpst atvasinātās nozīmēs lietoti vārdi (34), (35) un kuras marķētājs var interpretēt dažādi atkarībā no tā, vai vārdu (piemēros *acis* un *celt*) atvasināto nozīmi uztver kā pārnestu vai kā tiešu.

(33) *veikt izpēti, dot ziņu, laba slava*

(34) *labas acis* ‘laba redze’

(35) *celt traci* ‘radīt traci’

Sākotnēji metaforiskas nozīmes laika gaitā tēlainību mēdz zaudēt, bet šis process ir pakāpenisks un subjektīvi interpretējams (Dancygier, Sweetser 2014, 34–35). VVL klasifikācija tika balstīta uz „Tēzaurā” atrodamajām pārnesto (un tiešo) nozīmju norādēm, kas ļauj VVL iekļaut vai nu kolokācijās (bez nozīmes pārnese), vai frazeoloģismos (ar nozīmes pārnese), lai VVL piešķirtais marķējums nebūtu pretrunā ar „Tēzaura” datiem.

5. Vairākvārdu leksēmu novietojums „Tēzaura” šķirķļos

„Tēzaurā” daudzos šķirķļos saglabāts no „Latviešu literārās valodas vārdnīcas” mantotais VVL novietojums vārda šķirķlī atkarībā no tā, vai vārdu savienojums ir saistīts ar kādu konkrētu tajā lietotā vārda nozīmi, ar vairākām vārda nozīmēm vai vispār nav saistīts ar kādu konkrētu vārda nozīmi (Ceplītis 1972, 10, 16). Šis princips īstenojas, piem., atšķirīgā frazēmu un idiomu novietojumā. Frazēma *spēļu elle* ‘spēļu nams, spēļu namu kopums’ vārda *spēle* šķirķlī ir redzama pie pamatnozīmes, turpretim vārda *elle* šķirķlī tam jāatrodas

stabilu vārdu savienojumu sadaļā šķirkļa beigās. Savukārt idiomās *aitas galva* ‘neattapīgs, dumjš cilvēks’ vieta ir stabilu vārdu savienojumu sadaļā šķirkļu *aita* un *galva* beigās.

„Tēzaura” dati šai ziņā nav konsekventi, jo laika gaitā „Tēzaurā” ir iekļautas VVL no dažādām vārdnīcām, bieži vien automātiski pievienojot VVL attiecīgo vārdu šķirklīem bez sīkākas semantiskas analīzes. Šobrīd „Tēzaurā” ir šķirklī, kuros idiomās atrodamas pie noteiktām vārdu nozīmēm, lai gan tām būtu jāatrodas šķirkļa beigās, un otrādi – frazēmas atrodas šķirkļa beigās arī tad, kad tām būtu jāatrodas pie konkrētas nozīmes. Nākotnē ieviešami konsekventi principi VVL novietojumam šķirklī: 1) frazēmas jānovieto vai nu pie konkrētas nozīmes vai šķirkļa beigās atkarībā no tā, kura frazēmas komponenta šķirklis tas ir, 2) idiomām vienmēr jāatrodas šķirkļa beigās, 3) kolokācijām jāatrodas tikai pie noteiktām vārda nozīmēm, jo pēc „Tēzaura” klasifikācijā izmantotās definīcijas tajās visi vārdi ir lietoti kādā no savām tiešajām nozīmēm. Terminu un taksonu novietojums šķirklī būtu jāveido pēc tādiem pašiem principiem kā frazeoloģismu un kolokāciju novietojums – atkarībā no nozīmes pārnese.

Šāds novietojums šķirklī ir īstenojams tikai viennozīmīgām VVL vai tādām VVL, kurām visas nozīmes pieder pie vienas VVL grupas. Sastopamas arī tādas VVL, kurām viena nozīme ir tieša, bet otra – pārnesta. Tas attiecas uz frazeoloģismiem, kas ir nefrazeoloģisku vārdu savienojumu homonīmi, piem., VVL *gāzt no kājām* ir divas nozīmes – burtiskā ‘apgāzt, nogāzt cilvēku’ un idiomātiskā ‘iedarbojoties uz (kāda) organismu, traucēt tā normālu funkcionēšanu (piem., par alkoholiskiem dzērieniem, slimībām)’. Tiešā nozīme marķēta kā kolokācija, kurai vārdu *gāzt un kāja* šķirklī būtu jāatrodas pie attiecīgajām vārda nozīmēm, bet šobrīd VVL ar abām nozīmēm ir novietota šķirklī *gāzt un kāja* beigās.

Tehniski šādas nekonsekvences varētu risināt ar homonīmisku VVL ieviešanu, katru homonīmu kā atsevišķu lemmu novietojot vajadzīgajā vietā. Otrs risinājums varētu būt parādīt VVL vienā šķirklī divās vietās – gan pie vārda nozīmes, gan šķirkļa beigās – tomēr tas radītu nevēlamu informācijas atkārtoto. Labākais tehniskais risinājums vārdnīcas veidotājiem būs jāizvēlas nākotnē, bet raksta autore uzskata, ka VVL novietojums šķirklī ir vērtīgs paņēmieni, kā uzskatāmi parādīt, vai vārds šajā VVL ir saglabājis savu nozīmi.

Secinājumi un nākotnes ieceres

Rakstā aprakstītā „Tēzaura” VVL klasificēšana veikta lielākajai daļai no atbilstošā vārdnīcas materiāla (apm. 92 %). Empīriski pēc vārdnīcā esošā materiāla izveidotas klasifikācijas pamatgrupas – vairākvārdu vietvārdi, taksoni, termini, frazeoloģismi un kolokācijas. Papildu klasifikācijā tika šķirti divi latviešu valodas frazeoloģijā aprakstītie frazeoloģismi tipi – idiomās un frazēmas.

Darba gaitā tika konstatētas dažādas grūtības grupu nošķiršanā. Vismazāk grūtību bija taksonu jeb sugu nosaukumu nošķiršanā, tomēr nācās secināt, ka visa taksonu grupa nav terminu apakšgrupa, jo paralēli oficiālajiem nosaukumiem tiek lietoti arī dažādi neoficiāli augu un dzīvnieku nosaukumi. Lielākas grūtības bija terminu grupas noteikšanā, jo atsevišķām nozarēm nav oficiāli reģistrētu vai vārdnīcās publicētu

terminu, kā arī termini var būt tēlaini ar nozīmes pārnēsumu, kas parāda to pārklāšanos ar frazeoloģismu grupu. Vislielākās grūtības sagādāja frazeoloģismu apakštīpu nošķiršana un frazeoloģismu un kolokāciju nošķiršana, jo kritēriji par kāda VVL elementa nozīmes pārnēsumu un nozīmes saglabāšanu frazeoloģismā nav skaidri definēti attiecībā pret dažādām iespējamām nozīmes pārnēsuma pakāpēm, tāpēc frazeoloģismu dalījums idiomās un frazēmās jāpārvērtē un jāmeklē jauni kritēriji precīzākai to nošķiršanai, vai arī jāizvirza citāds dalījums.

Ieguvumi no VVL iedalīšanas mazākās grupās ir dažādi. Vārdnīcas lietotājiem informācija par VVL tipu palīdz labāk orientēties vārdnīcas materiālā, īpaši šķirkļos, kuros parādās liels skaits dažādu tipu VVL.

„Tēzaura” veidotājiem rodas iespēja iegūt priekšstatu par to, no kā sastāv VVL kopums, jo milzīgā datu apjoma dēļ visus vārdu savienojumu veidus ir gandrīz neiespējami pamanīt. Tiek iegūta informācija, cik VVL ir katrā no izveidotās klasifikācijas grupām, kas dod iespēju nepieciešamības gadījumā veidot sīkāku klasifikāciju, piem., taksonu grupu var dalīt sīkāk oficiālos nosaukumos un sarunvalodā lietotos nosaukumos. Kolokāciju grupa arī nav viendabīga, tāpēc tā nākotnē jāpēta un jādala apakštīpos, kā arī jāprecizē, kāda veida kolokācijas turpmāk iekļaut vārdnīcas materiālā. Pēc datu apkopošanas ir vieglāk risināma frazeoloģismu robežu noteikšana un variantu apvienošana, kas novērstu dublētu informāciju. Minētajās grupās neiekļautajām VVL nepieciešams ieviest jaunus tipus VVL klasifikācijā, lai tā aptvertu visu „Tēzaurā” esošo VVL materiālu. Tāpat arī nākotnē vārdkopterminu un taksonu marķējumu varētu papildināt ar informāciju par to tēlainību, t. i., katru no tiem varētu iekļaut arī frazēmu, idiomu vai kolokāciju grupā.

Iegūtā sākotnējā klasifikācija ir ieguvums valodniekiem, jo iespējams veikt katras grupas dziļāku izpēti. Piem., šķirošanas rezultātā iegūtā frazeoloģismu grupa ir bagātīgs materiāls pētījumiem par frazeoloģismu variēšanos un okazionāliem frazeoloģismiem.

Saīsinājumi

LNTP	<i>Latvijas Nacionālais terminoloģijas portāls</i>
VVL	vairākvārdu leksēma

Avoti

- Baldunčiks, Juris (red.). 1999. *Svešvārdu vārdnīca*. Rīga: Jumava.
- Latvijas Nacionālais terminoloģijas portāls*. Pieejams: <https://termini.gov.lv>
- Millers, Laimonis. 2007. *Jūrnecības vārdnīca*. Rīga: Zvaigzne ABC.
- Spektors, Andrejs *et al.* 2023. *Tēzaur.lv, 2023, Rudens laidieni*. CLARIN-LV digitālā bibliotēka. Pieejams: <http://hdl.handle.net/20.500.12574/92>

Literatūra

- Bauer, Laurie. 2022. *An Introduction to English Lexicology*. Edinburgh: Edinburgh University Press. <https://doi.org/10.1515/9781474477918>

- Burger, Harald. 2007. Semantic aspects of phrasemes. *Phraseologie. Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung. An International Handbook of Contemporary Research*. Burger, Harald, Dobrovolskij, Dmitrij, Kühn, Peter, Norrick, Neal R. (eds.). Berlin, New York: Walter de Gruyter, 90–109. <https://doi.org/10.1515/9783110197136>
- Ceplītis, Laimdots (atb. red.). 1972. *Latviešu literārās valodas vārdnīca*. 1. Rīga: Zinātne.
- Dancygier, Barbara, Sweetser, Eve. 2014. *Figurative Language*. Cambridge: Cambridge University Press. <https://doi.org/10.1075/etc.8.2.07rim>
- Grasmanis, Mikus *et al.* 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference*. Medved', Marek, Měchura, Michal, Kosem, Iztok, Kallas, Jelena, Tiberius, Carole, Jakubiček, Miloš, Krek, Simon (eds.). Brno: Lexical Computing CZ s.r.o., 400–418. Pieejams: <https://elex.link/ojs/index.php/elex/article/view/38>
- Matthews, Peter. 1997. *Concise Dictionary of Linguistics*. Oxford: Oxford University Press.
- Laua, Alīse. 1992. *Latviešu valodas frazeoloģija*. Rīga: Zvaigzne.
- Leseva, Svetlozara, Mititelu, Verginica Barbu, Stoyanova, Ivelina. 2020. It takes two to tango – towards a multilingual MWE resource. *Proceedings of the Fourth International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*. Koeva, Svetla, Osenova, Patya (eds.). Sofia: Bulgarian Academy of Sciences, 101–111. Pieejams: <https://aclanthology.org/2020.clib-1.0.pdf>
- Mel'čuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*. 3, 31–56. <https://doi.org/10.1515/phras-2012-0003>
- Skujiņa, Valentīna (red.). 2007. *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga: Latvijas Universitātes Latviešu valodas institūts.
- Spektors, Andrejs *et al.* 2016. Tezaurs.lv: the largest open lexical database for Latvian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Calzolari, Nicoletta *et al.* (eds.). Portorož: European Language Resources Association, 2568–2571. Pieejams: <https://aclanthology.org/L16-1408/>
- Škapare, Sanita. 2020. Ēdienkaršu valoda: ko un kā tās mums stāsta. *Valoda: nozīme un forma*. 11, 178–191. <https://doi.org/10.22364/vnf.11.11>
- Veisbergs, Andrejs. 2013a. Phraseological units and idioms, East and west and where do we stand. *Baltic Journal of English Language, Literature and Culture*. 3, 109–121. <https://doi.org/10.22364/BJELLC.03.2013.09>
- Veisbergs, Andrejs. 2013b. Latviešu valodas frazeoloģija. Veisbergs, Andrejs (red.). *Latviešu valoda*. Rīga: Latvijas Universitātes Akadēmiskais apgāds, 157–174.
- Veisbergs, Andrejs. 2019. The fuzzy concept of idiom and what it might mean for bilingual dictionaries. *Baltic Journal of English Language, Literature and Culture*. 9, 111–129. <https://doi.org/10.22364/BJELLC.09.2019.08>

Summary

„Tēzaurs” is the largest Latvian electronic dictionary that consists of more than 400,000 entries, including more than 71,000 multi-word expressions (MWEs) from a wide range of sources; these expressions are very varied in form and content. Over the recent years, there has been ongoing work of sorting the MWEs into several groups: phrasemes, idioms, collocations, complex terms, taxonomic group names and multi-word nouns. This article describes the current results of sorting the MWEs and the challenges associated with borderline cases or overlapping categories, as

well as unclear criteria for MWE division. It was concluded that additional criteria were needed to distinguish between phrasemes and idioms, as well as between phrasemes and collocations.

There are several advantages to dividing MWEs into smaller groups. For dictionary users, the additional information has improved the browsing of the dictionary contents. For the developers of „Tēzaurs”, this process has given more insight into the contents of the MWE data, which enables them to analyse an entire group of MWEs at once, prevent duplicates or discrepancies, amend their positioning in the dictionary entries, and improve the overall quality of the dictionary. For linguists, the newly assembled and structured language material allows for more in-depth studies of each MWE group and highlights new directions of research.

The current system of MWE classification is the first step towards an organised system of MWE description and classification, which will require further revision and improvement in the future.

Keywords: multi-word expressions; phraseology; collocations; multi-word terms; electronic dictionary „Tēzaurs”.



Rakstam ir Creative Commons Attiecinājuma 4.0 Starptautiskā licence (CC BY 4.0) / This editorial is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) (<https://creativecommons.org/licenses/by/4.0/>)