Latviešu valodas sintaktiskā korpusa atveidošana atbilstoši UD vadlīnijām
Atbilstības ar SemTi-Kamola vadlīnijām

Lauma Pretkalniņa
LUMII AILab

1. Dokumentā izmantotie apzīmējumi

UD — Universal Dependencies.
LVTB — Latvian Treebank.
A → B — A ir atkarības galva, B — atkarīgais (šāds bultas virziens ir gan TrEd, gan Brat).
Jautājums — neatrisināts jautājums, atbilde, kas jāiekļauj dokumentā.
Vārds, vārds, vārds — vārdu saraksti vietās, kur marķējumu piešķir vārdam no noteikta saraksta.
Teksts — UD ir paredzēts, bet esošais marķējums neļauj to iegūt.
Dokuments atjaunots, pamatdaļa rakstīta atbilstoši UD v2 vadlīnijām.

2. Dalīšana tekstvienībās

UD standarts paredz divu līmeņu dalīšanu tekstvienībās — vispirms sadala pa atstarpēm un pa pieturzīmēm, pēc tam, ja nepieciešams, šādi iegūtos vārdus var dalīt sīkāk, lai iegūtu *sintaktiskos vārdus* — angļu valodai sadala "king's" par "king" + "'s", kā arī sadala saīsinājumus "it's" un atdala klītikas (angl. *clitics*). Tas nenozīmē, ka fleksīvajās valodās vajadzētu dalīt vārdus pa morfoloģiskajām sastāvdaļām. *UD dokumentācija (pie pazīmes Reflex) iesaka atgriezeniskos darbības vārdus uzskatīt par elementu, kas jāsadala par darbības vārdu un ieplūdušu atgriezenisku vietniekvārdu, taču vismaz pagaidām šo realizēt mums nav pa spēkam.* **Nepieciešams regulāri sekot līdzi šim jautājumam, jo nostāja UD diskusijās vēl nav pilnīgi detalizēta.**

Citas viena vārda dalīšanas nepieciešamības ir saistītas ar labotajām kļūdām.

2.1. Tekstvienības, kurās ir atstarpes

UD v2 standarts šobrīd atļauj lietot *vārdus ar atstarpēm* specifiskos, skaidri definējamos gadījumos, taču neatļauj vārdus ar atstarpēm lietot, lai vairāku vārdu frāzes padarītu par vienu vārdu.

LVTB UD versija *vārdus ar atstarpēm* lieto šādos gadījumos:

- skaitļi un telefona numuri (xn un xx), piemēram, 2 358 000 un +371 20001234;
- saīsinājumi *u.t.jpr., u.c., u.tml., v.tml.*, *u.t.t.*, *N.B.*, *P.S.* un *P.P.S.* ar jebkuru *P.* daudzumu, ja tajos ir atstarpes.

Transformēšanas stratēģija:

- xf: sadalīt pa atstarpēm, marķēt kā unstruct, kas sastāv tikai no xf tipa elementiem;

- citi gadījumi — *lai arī*, *kaut gan* — tiek dalīti kā atsevišķas tekstvienības jau hibrīdmarķējumā.

2.2. Teksta redakcionālo labojumu atveide

Dalīšana tekstvienībās atšķiras šādos gadījumos.

1. Ja oriģināltekstā ir atsevišķu burtu kļūdas, tad FORM laukā liek vārdformu, kāda tā ir oriģināltekstā, norāda pareizo lemmu un tagu, FEATS kolonnā pievieno piezīmi Typo=Yes un MISC kolonnā norāda pareizo formu. *UD standarts prasītu šķirt nepareizi uzrakstītas vārdformas no nepareizi izvēlētām vārdformām, bet tas pašlaik nav iespējams.*
2. Ja vārds oriģināltekstā kļūdas dēļ ir rakstīts vairākos vārdos, piemēram, *tas ir ne vajadzīgs*, tad to attēlo kā vairākas ar *goeswith* savienotas tekstvienības. Pirmajam šobrīd tiek likts pareizais tags, lemma un MISC kolonnā tiek pievienota pareizā forma, pārējiem liek tagu X un lemmu nenorāda. Visiem, izņemot pēdējo, pieliek CorrectSpaceAfter=No. Vairākciparu skaitļus ar atstarpēm uzskata par vienu tekstvienību.
3. Ja oriģināltekstā ir bijis ievietots lieks komats, to ar *punct* lomu pakārto vārdam tieši pirms tā, UPOSTAG ir X, lemmu nenorāda, MISC kolonnā ieliek norādi par labojumu.
4. Ja oriģināltekstā trūkst komata, pirms tā esošā elementa MISC kolonnā ieliek norādi par šo komatu.
5. Ja oriģināltekstā trūkst atstarpes, kas ir vajadzīga, tad vārdam, aiz kura tai jābūt, MISC kolonnā pievieno norādi CorrectSpaceAfter=Yes.

   MISC kolonnā izmantotās norādes par labojumiem:

- CorrectionType — labojuma veids (standarts LVUDTB ietvaros)
  - Spelling — drukas kļūdas
  - Spacing — vārds bijis rakstīts ar nevajadzīgu atstarpi vai pārnesumu jaunā rindā pa vidu
  - InsertedPunctuation — ievietota oriģināltekstā neesoša pieturzīme
  - Inserted — ievietota tekstvienība, kas nav pieturzīme — vērtība, kas norāda uz problēmu oriģināldatos
  - RemovedPunctuation — izmesta oriģināltekstā lieki ievietota pieturzīme
- CorrectForm — pareizā vārdforma (UD standarts) [UD typos]
- CorrectSpaceAfter — pareizais atstarpjojums aiz tekstvienības (UD standarts) [UD typos]

3. Morfoloģija

UD morfoloģiskajam marķējumam ir 3 elementi — lemma, vārdšķira, pārējās pazīmes. Vārda vienkārša nosaukšana un vārda lietojums kontekstā morfoloģiski tiek marķēta vienādi.

3.1. Lemma

Saīsinājumus neatšifrē, cilvēku vārdus nevispārina ("Jāņa" lemma ir "Jānis", nevis abstrakts "Person name").

Pamatā lietojam tās lemmas, kas ir korpusā. **Izņēmumi:**

- noliegtiem verbiem v..[^p].{6}y.* un noliegtiem divdabjiem v..p.{8}y.* tiek ņemta nost negācija — kopš apmēram 2021. gada marta sākuma šis vairs formāli netiek darīts, jo jau pamatkorpuss satur nenoliegtas lemmas.
- kā kļūdu korekcijas mehānisms prievārdiem s.* un relatīvajiem apstākļa vārdiem rr.* lemmas tiek pārveidotas uz mazajiem burtiem, bet **šīm lemmām nemaz nevajadzētu būt ar lielajiem burtiem.**

3.2. Vārdšķiras

Problēmas ar DET: UD pieprasa šādu vārdšķiru izdalīt arī valodās, kurās par to parasti nerunā, piemēram, čehu. Tātad arī latviešu. Vispārējā vārdšķiru dokumentācija komentē, ka vietniekvārdi aizpilda lietvārdu pozīciju, un, lai gan dažās vietās valodās par vietniekvārdiem uzskata arī tos, kas aizpilda īpašības vārdu pozīciju, UD kontekstā tie uzskatāmi par DET.

Divdabjus UD liek ielikt pie VERB vai ADJ. Atkarībā no tā, kur liekam, droši vien atšķiras, kas ir saprātīga lemma.

Ģenitīveņus uzskata par lietvārdiem ģenitīvā. Ja izdodas atrast kādas paralēles citās valodās, var pārskatīt šo nostāju.

Daudz pārdomu ir radījuši palīgverbi. Versijās 2.1. līdz 2.7. izvēlējāmies kā palīgverbus AUX marķēt ne tikai *būt*, bet arī *kļūt*, *tikt* un *tapt* tālāk citēto iemeslu dēļ (paturu šo fragmentu dokumentā kā vēsturiski interesantu spriedumu:

> UD dokumentācija no vienas puses atļauj valodām izvēlēties, kā šķirt AUX/VERB, no otras puses pieprasa, lai lomas *aux* un *cop* tiktu lietotas tikai AUX vārdšķiras vārdiem.
> Apzinājām argumentus, ka (a) palīgnozīmes lietojumos *būt*, *kļūt*, *tikt*, *tapt* neveido ciešamo kārtu (nepilna formu sistēma), (b) *kļūt* pamatā lieto sastata izteicējos, *tikt* — saliktos laikos, bet *tapt* — mūsdienās gandrīz nekur (valoda attīstās uz lietojumu nepārklāšanos), (c) atšķirību, kāpēc cilvēki izvēlas lietot vienu vai otru, starp *būt* un *kļūt*/*tikt*/*tapt* var puslīdz iespiest TAME kategorijā "tense". Pateicoties tiem, nolēmām līdz tālākām UD dokumentācijas izmaiņām šos četrus uzskatīt par palīgdarbības vārdiem UD izpratnē, ja tie LVTB ir lietoti ne patstāvīgā nozīmē.

No v2.8. *kļūt* no AUX grupas izslēdzam pilnībā un *tikt*, *tapt* pieļaujam kā AUX tikai saliktajos laikos.

Vārdšķiru atbilsmes

Tags LVTB	**Vārdšķira, lemmu grupa LVTB**	**UD POSTAG**	**Piezīmes**
nc.*	Lietvārds: sugas	NOUN	
np.*	Lietvārds: īpašvārds	PROPN	
vc.*	Darbības vārds: *būt*, *nebūt* palīgdarbības vārda lietojumā	AUX	Attiecas gan uz divdabjiem, gan citām formām.
va.*	Darbības vārds: *tikt*, *netikt*, *tapt*, *netapt* saliktajos laikos	AUX	Attiecas gan uz divdabjiem, gan citām formām. Uz UD v2.8 ir pārskatīta izpratne, un šie vārdi saitiņas lietojumā (vt.*) vairs netiek uzskatīti par AUX.
v..[^p].*	Darbības vārds: citi	VERB	Fāzes un izpausmes veida verbus atstājam kā VERB, jo latviešu valodā pamata darbības vārdiem un dažādiem palīgdarbības vārdiem ir stipri līdzvērtīgi pilna formu sistēma, un vispār tos ir grūti nodalīt, nebalstoties dziļi semantiskos kritērijos.
v..pd.*	Lokāmais divdabis: citi	VERB	UD ļauj izvēlēties starp VERB/ADJ, paliekam pie VERB lemmas dēļ, turklāt ADJ disonētu ar lēmumu palīgverbu divdabjus marķēt kā AUX
v..pp.*	Daļēji lokāmais divdabis: citi	VERB	
v..pu	Nelokāmais divdabis: citi	VERB	UD ļauj izvēlēties starp VERB/ADV, paliekam pie VERB lemmas dēļ, turklāt ADV disonētu ar lēmumu palīgverbu divdabjus marķēt kā AUX
a.*	Īpašības vārds: *manējais, tavējais, savējais, mūsējais, jūsējais, viņējais*	DET/PRON	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON. Pēc līdzības ar čehu piemēru *svůj versus jeho, její, jejich* sk. http://universaldependencies.org/cs/pos/DET.html **Saprast, kā šis ir domāts par PRON** **http://universaldependencies.org/u/overview/morphology.html#pronominal-words** **Nepieciešams papildināt valodspecifisko aprakstu UD lapā, uzskaitot visus DET.**
a.*	Īpašības vārds: *daudzi, vairāki*	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON. *Daži* un *citi* jau tiek marķēti kā vietniekvārdi.
a.*	Īpašības vārds: citi	ADJ	
pp.*	Vietniekvārds: personas	PRON	
px.*	Vietniekvārds: atgriezeniskais	PRON	
ps.*	Vietniekvārds: piederības	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON. Pēc līdzības ar čehu piemēru *svůj versus jeho, její,*

			jejich sk. http://universaldependencies.org/cs/pos/DET.html .
pd.*	Vietniekvārds: norādāmais	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON.
pi.*	Vietniekvārds: nenoteiktais	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON.
pq.*	Vietniekvārds: jautājamais	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON.
pr.*	Vietniekvārds: attieksmes	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON.
pg.*	Vietniekvārds: noteiktais	PRON -> PRON/DET	Ja nav pieejams sintakses koks, tad PRON. Ja UD sintakses koks ir pieejams un tajā loma ir *det*, tad DET, citādi PRON.
r0.*	Apstākļa vārds: **vietniekvārdisks**	ADV*/SCONJ*	*Šī brīža marķējums neļauj izšķirt apstākļa vārdus, kas ievada palīgteikumus.*
r[^0].*	Apstākļa vārds: citi	ADV	
mc.*	Skaitļa vārds: pamata	NUM	
mo.*	Skaitļa vārds: kārtas	ADJ	
mf.*	Skaitļa vārds: daļskaitlis	NUM	
s	Prievārds	ADP	
cc.*	Saiklis: sakārtojuma	CCONJ	
cs.*	Saiklis: pakārtojuma	SCONJ	
i	Izsauksmes vārds	INTJ	
q	Partikula	PART	UD dokumentācija atzīst, ka PART ir tas, kas neiederas citur. **Vajag UD dokumentācijā dot partikulu uzskaitījumu.**
z.*	Pieturzīme	PUNCT	**TODO: UD ir izskanējis apgalvojums, ka pieturzīmēm vajadzētu būt lomā *punct* un dažādi jocīgi reziduāļi, kas nokļūst šajā lomā, nedrīkst būt SYM. Jāpārbauda, kas mums sanāk.**
yn	Saīsinājums: sugasvārdisks	NOUN	
yp	Saīsinājums: īpašvārdisks	PROPN	
ya	Saīsinājums: adjektīvisks	ADJ	
yv	Saīsinājums: verbāls	VERB	Pieņemam, ka AUX tipa darbības vārdus nesaīsina
yr	Saīsinājums: adverbiāls	ADV	

yd	Saīsinājums: diskursa iezīmētājs	SYM	**Diskusija**: Nez, kā tas ir pareizi domāts? Angļi etc. marķē ar X (http://universaldependencies.org/en/pos/X.html), franči etc. marķē ar SYM, zviedri kategoriju X vispār nelieto. Dažādās dokumentācijās (piemēram, http://universaldependencies.org/it/pos/all.html#al-it-pos/X) slēpti uzpeld attieksme, ka X ir tas, kam nav nekādas jēgas un saturīgas interpretācijas — šādi skatoties utt. noteikti nav X, angļi vienkārši nav pārmarķējuši.
xf	Reziduālis: svešvalodā	X*/PROPN/ NOUN*	*Šī brīža marķējums vismaz daļā gadījumu neļauj izšķirt reziduāļus, kas lietoti kā lietvārdi, piemēram: PROPN: Hennessy, [M.] Fisher Boel, [P.] Cox, Lancome, [K.] Klemm u.c (pēc pirmā burta atšķirt nevar, piemēram, vācu valodas tekstos).* *NOUN: matique naturell, art deco u.c.*, **vai ir vienvārdīgi piemēri?** *X: no such nick..; primary education u.c.* *vēl tā marķēti: FUCK; fon*
xn	Reziduālis: skaitlis ciparіem	NUM	
xo	Reziduālis: kārtas skaitlis ciparіem	ADJ	
xu	Reziduālis: URL	PROPN	Avots: https://github.com/UniversalDependencies/docs/issues/973#issuecomment-1859447774
xx	Reziduālis: Cits	SYM*/PROPN/ NOUN*	*Šī brīža marķējums neļauj izšķirt reziduāļus, kas lietoti kā lietvārdi.*

3.3. Pazīmes

Kādas UD pazīmes var no mūsu tagiem izgūt.

Jāņem vērā, ka šīs pazīmes tiek piešķirtas vienai tekstvienībai, tāpēc te neapraksta analītiskās formas.

3.3.1. Inflectional features: nominal

- Gender — mūsu tagsetā lokāmiem lietvārdiem, īpašības vārdiem, vietniekvārdiem, dažiem divdabjiem (lokāmajiem, daļēji lokāmajiem), dažiem skaitļa vārdiem.
    - Masc (masculine gender): n.m.*, a.m.*, v..p.m.*, p..m.*, m..m.*
    - Fem (feminine gender): n.f.*, a.f.*, v..p.f.*, p..f.*, m..f.*
    - Neut (neuter gender), Com (common gender) — nav un nevajag, jo kontekstā vienmēr visi, kam dzimte vispār var piemist, ir sieviešu vai vīriešu dzimtē.
- Animacy — nav un nevajag.

- Number — mūsu tagsetā lokāmiem lietvārdiem, darbības vārdiem un dažiem divdabjiem, īpašības vārdiem, dažiem vietniekvārdiem, skaitļa vārdiem
    - Sing (singular number): n..s.*, v..[^p]….s.*, v..p..s.*, a..s.*, p…s.*, m…s.*
    - Plur (plural number): n..p.*, v..[^p]….p.*, v..p..p.*, a..p.*, p…p.*, m…p.*
    - Dual (dual number): tagsetā nav, korpusā nav, Tēzaurā bija laikam viens piemērs, bet nevajag, jo ir praktiski atmiris.
    - Count (count plural) — it kā nav. **Noskaidrot, vai varētu būt, ka Number=Count ir attiecināms uz kaut kādiem mūsu daudzuma ģenitīva gadījumiem?!**
    - Ptan (plurale tantum) — n..d.* (bikses, durvis u.c. daudzskaitlinieki) — ja nav atzīmēts, tad nav, rēķināmies, ka robeža var būt izplūdusi.
    - Coll (collective / mass / singulare tantum) — n..v.* — vienskaitlinieki — to, kas jau ir atzīmēts, varam turpināt atzīmēt, rēķināmies, ka vienskaitlinieki valodā ļoti strauji kļūst par normāliem lietvārdiem, tāpēc robeža ir vēl vairāk izplūdusi nekā daudzskaitliniekiem.
    - Tri (trial number), Pauc (paucal number), Grpa (greater paucal number), Grpl (greater plural number), Inv (inverse number) — nav.
- Case — mūsu tagsetā lietvārdiem, īpašības vārdiem, dažiem divdabjiem, vietniekvārdiem, skaitļa vārdiem
    - Nom (nominative / direct) — nominatīvs — n…n.*, a…n.*, v..p…n.*, p….n.*, m….n.*
    - Acc (accusative / oblique) — akuzatīvs — n…a.*, a…a.*, v..p…a.*, p….a.*, m….a.*
    - Dat (dative) — datīvs — n…d.*, a…d.*, v..p…d.*, p….d.*, m….d.*
    - Gen (genitive) — ģenitīvs — n…g.* **(TODO n…gg.*?)**, a…g.*, v..p…g.*, p….g.* m….g.*
    - Voc (vocative) — vokatīvs, ir atsevišķi piemēri atzīmēti — n…v.*, a…v.*, v..p…v.*
    - Loc (locative) — lokatīvs — n…l.*, a…l.*, v..p…l.*, p….l.*, m….l.*
    - Ins (instrumental / instructive) — bezprievārda instrumentāļi netiek marķēti, valodā reti sastopami, pamatā vecos tekstos.
    - Abs (absolutive), Erg (ergative), Par (partitive), Dis (distributive), Ess (essive / prolative), Tra (translative / factive), Com (comitative / associative), Abe (abessive), Ine: (inessive), Ill (illative), Ela (elative), Add (additive), Ade (adessive), All (allative), Abl (ablative), Sup (superessive), Sub (sublative), Del (delative), Lat (lative / directional allative), Tem (temporal), Ter (terminative / terminal allative), Cau (causative / motivative), Ben (benefactive / destinative), Cmp (comparative), Equ (equative) — nav.

- Definite (definitnes or state) — īpašības vārdiem, dažiem divdabjiem; tā kā UD kārtas skaitļa vārdus uzskata par īpašības vārdiem, tiem norāda noteiktību (visiem Def)
    - Ind (indefinite) — a…..n.*, v..p……n.*, kā arī trešs, ceturts, utt.
    - Spec (specific indefinite) —nav, bet vajag **noskaidrot, varbūt te tomēr ir trešs un ceturts? Ja ir, kā atļaut?**
    - Def (definite) — a…..y.*, v..p……y.*, mo.* (izņemot "trešs", "ceturts")
    - Cons (construct state), Com (complex) — nav un nevajag.
- Degree — īpašības vārdiem, daļai apstākļa vārdu, dažiem divdabjiem.

  Tādas patoloģijas kā "pirmais, pirmākais, vispirmākais" pašlaik paliek neapskatītas.
    - Pos (positive, first degree) — a…..p.*, rp.*, v..pd……p.*, kārtas skaitļa vārdi mo.*
    - Cmp (comparative, second degree) — a…..c.*, rc.*, v..pd……c.*
    - Sup (superlative, third degree) — a…..s.*, rs.*, v..pd……s.*
    - Equ (equative), Abs (absolute superlative) — nav

3.3.2. Inflectional features: verbal

Mulsinoši, ka mums sanāk tik maz raksturlielumu daļēji lokāmajam divdabim un abiem nelokāmajiem. Kas mums atšķir abus nelokāmos divdabjus vienu no otra?

- VerbForm — var piemist ne tikai darbības vārdiem, bet arī citām vārdšķirām, piemēram, lokāmajiem divdabjiem, kas marķēti kā ADJ, vēlams likt VerbForm=Part
    - Fin (finite verb) — ieteikums no [ud-web-morfo]: ja kaut kam ir Mood, tad tas ir arī Fin, tāpēc nesakrīt ar LV tradicionālo izpratni, kurā finītās formas ir tās, kurām ir personas kategorija. Vienojamies UD transformācijai izmantot UD izpratni un par finītām marķēt visas verbu formas, izņemot divdabjus un nenoteiksmi v..[^pn].*
    - Inf (infinitive) — v..n.*
    - Sup (supine) — vecos dialektu tekstos pa retam ir sastopams, mūsdienās atmiris — *nāc ēstu, aizgāja pienu dzertu*.
    - Part (participle, verbal adjective) — darbības vārda un īpašības vārda hibrīds — v..pd.*, īpašības vārdi, kam lemma beidzas ar -ošs, -oša (**pārbaudīt datos**), *daži citi īpašības vārdi*

      Šeit vajadzētu iekļaut arī adjektivizējušos lokāmos divdabjus, kas LVTB jau ir nomarķēti kā īpašības vārdi, bet kā tādus atšķirt?
    - Conv (converb, transgressive, adverbal participle, verbal adverb) — darbības vārda un apstākļa vārda hibrīds — v..pu.*, v..pp.*, (nelokāmos un daļēji lokāmos divdabjus iekļaujam šeit, jo citur tie neiederas vēl vairāk), *daži apstākļa vārdi*

Šeit varētu iekļaut tādus vārdus kā "ziedošāk" — apstākļi, kas veidoti no lokāmajiem divdabjiem, taču šāds marķējums šobrīd nav pieejams.

  - Gdv (gerundive) — nav
  - Ger (gerund) — novecojusi vērtība, to neiesaka lietot, ja cita atbilst labāk. — nav.
  - Vnoun (verbal noun, masdar) — darbības vārda un lietvārda hibrīds — 4. dekl. lietvārdi n.....4.* ar lemmas izskaņu -šana un atgriezeniskie lietvārdi n.....r.* ar lemmas izskaņu -šanās, **kas vēl?**
- Mood — izteiksme — darbības vārdiem.
  - Ind (indicative) — īstenības izteiksme — v..i.*
  - Imp (imperative) — pavēles izteiksme — v..m.*
  - Cnd (conditional) — vēlējuma izteiksme — v..c.*
  - Pot (potential) — nav
  - Sub (subjunctive / conjunctive) — nav
  - Jus (jussive) — nav
  - Prp (purposive) — nav
  - Qot (quotative) — atstāstījuma izteiksme — v..r.*
  - Opt (optative) — nav
  - Des (desiderative) — nav
  - Nec (necessitative) — vajadzības izteiksme —v..d.*
  - Adm (admirative) — nav.
- Tense
  - Past (past tense / preterite / aorist) — vienkāršā pagātne finītajām formām v..[^p]s.*, lokāmie pagātnes divdabji v..pd….s.*
  - Pres (present tense) — vienkāršā tagadne finītajām formām v..[^p]p.*, lokāmie tagadnes divdabji v..pd….p.*
  - Fut (future tense) — vienkāršā nākotne v..[^p]f.*
  - Imp (imperfect), Pqp (pluperfect) — nav, jo tie ir speciāli pagātnes paveidi.
- Aspect — pabeigtība — latviešu valodā lielākoties izsaka ar leksiskiem paņēmieniem, robežas ir pietiekami neskaidras, tādēļ tās nav vērts marķēt.
  - Imp (imperfect aspect) — v..pd…ap.*(-ošs divdabji)
  - Perf (perfect aspect) — v..pd….s.* (-is, -ts divdabji)
  - Prosp (prospective aspect), Prog (progressive aspect), Hab (habitual aspect)
  - Iter (iterative / frequentive aspect) — nav, jo latviski iteratīva darbība izpaužas pamatā ar piedēkli (kāpt/kāpelēt), bet, iteratīvam darbības vārdam pieliekot

priedēkli, iterativitāte bieži atkal pazūd (raustīt/paraustīt/apraustīt/saraustīt) — nav nepārprotamu morfoloģisko pazīmju, pēc kurām skaidri noteikt

- Voice
    - Act (active voice) — darāmā kārta — finītajām formām v..[^p]…..a.* un lokāmajiem divdabjiem v..p…..a.*, *daži īpašības vārdi*.

      Šeit vajadzētu iekļaut arī adjektivizējušos lokāmos divdabjus, kas LVTB jau ir nomarķēti kā īpašības vārdi, bet kā tādus atšķirt?
    - Mid (middle voice) — pamatā nav (Holvūts (2001, 188) par vidējo kārtu latviešu val. ierosina saukt tādus dīvaiņus kā *pudele izdzērās*, *puķes nopirkās*, bet mums nav morfokritēriju tādu nošķiršanai)
    - Pass (passive voice) — ciešamā kārta — finītajām formām v..[^p]…..p.* un lokāmajiem divdabjiem v..p…..a.*, *daži īpašības vārdi*.

      Šeit vajadzētu iekļaut arī adjektivizējušos lokāmos divdabjus, kas LVTB jau ir nomarķēti kā īpašības vārdi, bet kā tādus atšķirt?
    - Antip (antipassive voice), Dir (direct voice), Inv (inverse voice), Rcp (reciprocal voice), Cau (causative voice) — nav
- Evident (evidentiality)— morfoloģiski marķēta norāde uz runātāja informācijas avotu, latviešu valodā — verbiem. **Vai tas ir pareizi, ka mums sanāk, ka dublējas ar Mood?**
    - Fh (first hand) — īstenības izteiksme, **kas vēl?**
    - Nfh (non-firsthand) — atstāstījuma izteiksme
- Polarity — darbības vārdiem un divdabjiem ņem no taga.

  UD dokumentācijā ir pieminēts, ka čehi to lieto arī priekš "balts, nebalts". Ja "nebalts" ir Neg, tad "balts" ir Pos, jo var pielikt "ne". Arī latviešu valodai negatīvu plaši veido arī pārējām vārdšķirām, bet tur ir grūti nošķirt bez marķējuma gan tehniski (ne- var būt gan priedēklis, gan saknes sākums), gan arī semantiski (kad šī ir "tikai" dotā vārda negācija un kad jauns vārds ar savu nozīmi, piemēram, darbs un nedarbs), tāpēc vismaz šobrīd plašāku marķējumu nevaram realizēt.
    - Pos (positive, affirmative) —v..[^p]……y.*, v..p……..y.*, q ar lemmu *jā*
    - Neg (negative) — v..[^p]……n.*, v..p……..n.* q ar lemmām *ne*, *nē*, cc ar lemmu *ne.*
- Person — vietniekvārdiem (arī tiem, kas kļūst par DET), darbības vārdiem, īpašības vārdiem *manējais, tavējais, jūsējais, mūsējais, viņējais*.
    - 0 (zero person) — nav, **varbūt "kas" un tamlīdzīgi?**
    - 1 (first person) — 1. persona vietniekvārdiem p.1.*, darbības vārdiem v..[^p]…1.*, kā arī *manējais, mūsējais*

- 2 (second person) — 2. persona vietniekvārdiem p.2.*, darbības vārdiem v..[^p]…2.*, kā arī *tavējais, jūsējais*
- 3 (third person) — 3. persona vietniekvārdiem p.3.*, darbības vārdiem v..[^p]…3.*, kā arī *viņējais*
- 4 (fourth person) — nav

- Polite — morfoloģiski marķēts pieklājības līmenis.

  Liekas, ka šobrīd nav veidu, kā nošķirt.
  - Infm (informal register) — **parastais tu/jūs lietojums?**
  - Form (formal register) — "jūs" lietots vienskaitlī un/vai ar lielo burtu, **kā lai atšķir (saliktajos laikos verbs vsk.: jūs esat gājis)?**
  - Elev (referent elevating), Humb (speaker humbling) — nav.

3.3.3. Lexical features

- PronType — vietniekvārdīguma veids, norādīts vietniekvārdiem un pronominālas cilmes adverbiem.
  - Prs (personal or possessive personal pronoun or determiner) — pp.*, ps.*, kā arī īpašības vārdi manējais, tavējais, mūsējais, jūsējais, viņējais, savējais, px.*.
  - Rcp (reciprocal pronoun) — nav.
  - Art (article) — determinētājs, kas ir noteiktības/nenoteiktības pazīme — nav.
  - Int (interrogative pronoun, determiner, numeral or adverb) — pq.*, daži apstākļa vārdi r0*.*: *cik, kad, kā, kur, kurp, kāpēc, kādēļ, kālab*, *kālabad* lietvārds n.* *kuriene* prepozicionālā savienojumā (*no kurienes, uz kurieni*).
  - Rel (relative pronoun, determiner, numeral or adverb) — pr.*, apstākļa vārdu nav (rr.* ir kaut kas cits).
  - Exc (exclamative determiner) — vietniekvārds, kas ievada izsaucienu vai kā citādi uzsver — piemēram, "kas par desām". *Šeit nav skaidrs, kā atšķirt no jautājamajiem vietniekvārdiem, ko UD arī atzīst kā tipisku problēmu daudzos tagsetos.*
  - Dem (demonstrative pronoun, determiner, numeral or adverb) — pd.*, daži apstākļa vārdi r0.*: *te, tur, šeit*, *tad, tagad, tik, tā*, lietvārdi n.* *turiene, tejiene* prepozicionālā savienojumā
  - Emp (emphatic determiner) — uzver nomināli, no kura ir atkarīgi — *pats* savienojumā *viņš pats* utml. **Kā lai nošķir? Kas te vēl var būt bez pats?**
  - Tot (total (collective) pronoun, determiner or adverb) — pg.*, daži apstākļa vārdi r0.* *vienmēr, visur*, *visad*, *visadiņ* lietvārdi n.* *visuriene*, *visadiene* prepozicionālā savienojumā.

- Neg (negative pronoun, determiner or adverb) — p…..y, apstākļa vārdi r0.*, kas sākas ar ne-, piemēram, *nekur, nekad*, lietvārdi n.* *nekuriene, nekadiene* prepozicionālā savienojumā.
- Ind (indefinite pronoun, determiner, numeral or adverb) — pi.*, apstākļa vārdi r0.*, lietvārds n.* *kuriene*, kas lietoti konstrukcijās ar partikulām *kaut, diez, diezin, nez, nezin*.

- NumType — skaitlīguma veids
  - Card (cardinal number or corresponding interrogative / relative / indefinite / demonstrative word) — mc.*, xn.*

    Lietvārdi, kas apzīmē skaitu, šobrīd netiek marķēti — desmits, simts, ducis, miljards.
  - Ord (ordinal number or corresponding interrogative / relative / indefinite / demonstrative word) — mo.*, xo.*, **vai kaut kas arī no citām grupām?**
  - Mult (multiplicative numeral or corresponding interrogative / relative / indefinite / demonstrative word) — r0.* ar lemmu *vienreiz, divreiz, trīsreiz, četreiz, piecreiz, sešreiz, septiņreiz, astoņreiz, deviņreiz, desmitreiz, pusotrreiz*.

    Citas, retākas "reizes" šobrīd ir izlaistas.
  - Frac (fraction) —mf.*

    Lietvārdi, kas apzīmē daļas nosaukumu, šobrīd netiek marķēti — puse, trešdaļa, desmitdaļa.
  - Sets (number of sets of things), Dist (distributive numeral), Range (range of values) — nav
- Poss (possessive)
  - Yes (it is possessive) — ps.*, kā arī īpašības vārdi manējais, tavējais, savējais, mūsējais, jūsējais, viņējais.
- Reflex (reflexive)
  - Yes (it is reflexive) — px.*, v.y.*

    Vismaz pirmajā versijā mums nav iespēju sadalīt atsevišķi katru atgriezenisko darbības vārdu par verbu un "atgriezenisko daļiņu", tāpēc šo pazīmi liekam arī atgriezeniskajiem verbiem.
- Foreign — vārds citā valodā
  - Yes (it is foreign) — xf.*
- Abbr (abbreviation) — saīsinājums
  - Yes (it is abbreviation) — y.*

4. Sintakse

4.1. Kas kam atbilst?

Atbilsme jādefinē divos aspektos — (1) kuras lomas kurās situācijās atbilst kurām un (2) strukturālās atbilsmes (gribētos jau 1:1, bet tik labi jau nebūs).

Kopā ar šo nodaļu ieteicams aplūkot diagrammu, kas shematiski parāda šobrīd vēl daļējās lomu atbilsmes. Diagramma pieejama: http://sintakse.korpuss.lv/docs/v2-14/LV2UD_mapping.pdf

4.1.1. Strukturālās atbilsmes

Pamatu dod [Nivre, et al, LREC] "Each word depends either on another word in the sentence or on a notional "root" of the sentence, following three principles: content words are related by dependency relations; function words attach to the content word they further specify; and punctuation attaches to the head of the phrase or clause in which it appears":

1. PMC
    a. Komati un saikļi tiek pakārtoti centrālajam darbības vārdam;
    b. Uzrunas un savrupinājumi tiek pakārtoti centrālajam darbības vārdam.
2. X-vārdi

    Ja viens no vārdiem ir pilnnozīmes vārds, tas kļūst par sakni virs funkcionālā vārda:

    a. xPrep, xParticle — basElem → prep, no
    b. xSimile — basElem → conj
    c. xPred —

        sastata izteicējs ar saitiņām *(ne)būt*, ~~*(ne)kļūt,*~~ *(ne)tikt*, *(ne)tapt*: basElem → visi auxVerb, pēdējais ir cop, pārējie ir aux, piemēram, *viņš ir bijis skolotājs — ir* = aux, *bijis* = cop, *skolotājs* = sakne, viss zem saknes;

        sastata izteicējs ar citu saitiņu: auxVerb → basElem, piemēram, *durvis ir stāvējušas atvērtas — stāvējušas* = sakne, *ir* = aux, *atvērtas* = xcomp);

        salikts laiks ar palīgdarbības vārdiem *(ne)būt*, ~~*(ne)kļūt,*~~ *(ne)tikt*, *(ne)tapt*: basElem → visi auxVerb;

        salikts laiks ar citu palīgdarbības vārdu: auxVerb → basElem;

        izteicējs ar modificētāju: mod (pats var būt salikts) → basElem (pats var būt salikts), piemēram, *viņš ir gribējis ēst — gribējis* = sakne, *ēst* = xcomp, *ir* = cop;, *viņš gribēja būt skolotājs — gribēja*= sakne, *skolotājs* = xcomp, būt = cop zem *skolotājs*.

        Kopš UD v2.8 *kļūt* tiek uzskatīts par pamata darbības vārda analogu, nevis par saitiņu/AUX.
    d. xNum — pirmie basElem ← pēdējais basElem
    e. xApp — pirmais basElem ← otrais basElem (mūsu datos 3+ nav bijuši)

f. xFunctor — pirmais basElem → pēdējie basElem

g. phrasElem, unstruct — pirmais basElem → otrais basElem; **kā velk atkarības, ja ir 3+ elementi?**

h. namedEnt — pirmais basElem → otrais basElem, pirmais basElem → trešais basElem, utt.

i. subrAnal — vadoties pēc tagā norādītā tipa

   vv: pirmais basElem → pēdējie basElem; **vai tā ir pareizi, ja ir 3+ elementi?**

   ipv: pārējie basElem ← pēdējais basElem ar tagu a.* vai ya.*

   skv: pirmais basElem ar tagu p.* ← pārējie basElem

   set: basElem, kas nav xPrep → basElem, kas ir xPrep

   sal: basElem, kas nav xSimile → basElem, kas ir xSimile

   ~~part: pirmais basElem → pēdējie basElem~~

j. unstruct — pirmais basElem → pēdējie basElem

k. **coordAnal — ?**

   Jāskatās atbilstošie ieteikumi dokumentācijā.

   UD mailinglistē Nivre atgādina, ka “all “name” and “mwe” structures are left-headed”.

3. Koordinācijas

   a. Vienlīdzīgi teikuma locekļi

      Pirmais vienlīdzīgais teikuma loceklis ir frāzes sakne, pārējie vienlīdzīgie teikuma locekļi pakārtoti pirmajam.

      Komati un saikļi pakārtoti tam vienlīdzīgajam teikumu loceklim, kas pirmais seko aiz attiecīgā komata vai saikļa.

      Ja aiz pēdējā vienlīdzīgā locekļa ir kādi saikļi, tos pakārto pēdējam vienlīdzīgajam loceklim. **Vai labāk pirmajam?**

   b. Vienlīdzīgas teikuma daļas, ko atdala ar komatu, analizē kā vienlīdzīgus izteicējus.

      Vienlīdzīgas teikuma daļas, ko atdala ar semikolu — paratakse. Semikolu nolēmām pakārtot otrās daļas izteicējam (UD viedoklis nav viennozīmīgi skaidrs).

      Ko dara, ja vairākas daļas, bet viens semikols?

      Kam ir pakārtots semikols — daļai pirms tā vai daļai pēc tā?

4. Virsotnes (situantiem, determinantiem, SPK), kas pakārtotas teikuma daļām — zem atbilstošā elementa saknes.

5. Redukcija — viens no atkarīgajiem tiek pacelts izlaistā vārda vietā, sīkāk sk. 4.3. nodaļu.

4.2. Komentāri par specifiskām lomām

4.2.1. *acl*: clausal modifier of noun (adjectival clause)

—

4.2.2. *advcl*: adverbial clause modifier

—

4.2.3. *advmod*: adverbial modifier

—

4.2.4. *amod*: adjectival modifier

—

4.2.5. *appos*: appositional modifier

—

4.2.6. *aux*: auxiliary

Apakšgadījums — aux:pass (passive auxiliary) v1 *auxpass* vietā. UD v2 *aux* var atzīmēt arī kaut kādas laika, personas partikulas, kādu latviešu valodā nav. UD v2 tiek ieteikts (**prasīts?**), lai *aux* lomu izpildītu AUX, nevis VERB.

4.2.7. *case*: case marking

—

4.2.8. *cc*: coordinating conjunction

—

4.2.9. *ccomp*: clausal complement

Klauzāls verba vai adjektīva atkarīgais, kam var būt pašam savs subjekts (gadījumos ja subjektu viennozīmīgi nosaka vecāks, tad jālieto *xcomp*). Šo lomu lieto tikai pēc valences obligātajiem paplašinātājiem. **Nepieciešams pieskatīt, lai mēs nesamarķējam neobligātās lomas šādi. Vai, ja to nevaram, tad vismaz godīgi atrunāt.**

***ccomp* lietojot arī ar saitiņām — *the important thing is to keep calm* un *the problem is that this has never been tried* — vai tas nozīmē, ka daļa izteicēja palīgteikumu ir *ccomp*? Nez, kāda loma otrajā teikumā viņiem ir *that*?**

Šo lomu nelieto lietvārda atkarīgajiem (tiem ir *acl*?)

Kopš 2.10 šo lomu lieto tiešajām runām (dirSp), pēc aprakstiem ir skaidrs, ka tas ir pareizais lietojums, ja tiešā runa aizpilda runas verba valenci. **Nav skaidrs, vai šis ir pieņemami teikumos "viņš paraustīja plecus: neko nezinu", kas šobrīd tiek marķēts tāpat.**

4.2.10. *clf*: classifier

Liekas, ka latviešu valodā nav nepieciešama.

4.2.11. *compound*: compound

Vai atbilstoši v2 ir pieņemami šo lietot vairākvārdu skaitļiem, vārdrindu un vārdkopu analogiem? "Nominal compounding" ir minēts dokumentācijā. Vēl UD v2 ir paredzēts *compound:prt* partikulverbiem, *compound:lvc* priekš "light verb constructions" un *compound:svc* priekš "serial verbs" — vairums šo situāciju latviešu valodā ir vienkārši priedēkļverbi. **Vai mūsu xPred kādos gadījumos var būt "serial verbs"**?

4.2.12. *conj*: conjunct

—

4.2.13. *cop*: copula

Izteicējos ar *cop* teikuma sakne ir izteicēja pamatdaļa. UD v2 tiek ieteikts (**prasīts?**), lai *cop* lomu izpildītu AUX, nevis VERB. **Nepieciešams pārbaudīt, ka tā ir realizēts.**

4.2.14. *csubj*: clausal subject

Apakšgadījums — *csubj:pass* (clausal passive subject) v1 *csubjpass* vietā. **Nepieciešams izanalizēt, vai mums pasīva šķīrums ir morfoloģiski pamatots?**

4.2.15. *dep*: unspecified dependency

LVTB nav šobrīd nepieciešams, bet var noderēt, apstrādājot parsētāju rezultātus, piemēram, kad parsētājs dod nepilnīgu rezultātu vai kad LVTB lomu nav iespējams transformēt.

4.2.16. *det*: determiner

—

4.2.17. *discourse*: discourse element

—

4.2.18. *dislocated*: dislocated elements

Latviešu valodā lielākoties neveidojas brīvās vārdu secības dēļ. Retos gadījumos (vārdu secība ir tik kliba, ka teikums kļūst negramatisks) šādu lomu lietot būtu noderīgi, taču mums LVTB šobrīd nav šādu teikumu.

4.2.19. *expl*: expletive

Latviešu valodā nav, taču, *ja atbilstoši UD vadlīnijām no atgriezeniskajiem darbības vārdiem tiktu izdalīta "atgriezeniskā daļiņa", tad to pamata darbības vārdam pievienotu ar šādu lomu.*

Sintaktiskie ekspletīvi, kas aizpilda obligātās lomas, ja tām nav saturiska aizpildījuma (*It is raining*) un klītikas atsevišķos gadījumos. Tiek nolemts, ka palīgteikumu atbalsta vārdi nav attiecināmi uz šo lomu, jo tie gandrīz nekad nav sintaktiski obligāti.

4.2.20. *fixed*: fixed multiword expression

Paredzēts fiksētām vienībām — funkcionālajiem vārdiem un īsiem adverbiem. **Nepieciešams izanalizēt, vai ir pieņemami šādi marķēt xSimile.**

4.2.21. *flat*: flat multiword expression

Daļēji fiksētām vairākvārdu vienībām. **Vai ir pieņemami šādi marķēt phrasElem un unstruct**?

Apakšgadījums — *flat:foreign* — svešvārdu virknēm (v1 vienkārši *foreign*)

Apakšgadījums — flat:name — bezstruktūras vārdiem (v1 vienkārši *name*)

4.2.22. goeswith: goes with

Nepieciešams realizēt, ka šādi marķē vārdus, kas kļūdas dēļ sadalīti divos vārdos, piemēram, *ne varēt*.

4.2.23. *iobj*: indirect object

—

4.2.24. *list*: list

LVTB nav marķēts, lai gan juridiskajos dokumentos šādi saraksti noteikti ir.

Kā diferencēt sarakstus no vienlīdzīgiem teikuma locekļiem?

4.2.25. *mark*: marker

—

4.2.26. *nmod*: nominal modifier

Nomināli paplašinātāji, kas modificē nomenu. **Kaut kad jāpapēta sīkāk, kā tur viss ir saistīts ar valenci.**

4.2.27. *nsubj*: nominal subject

Apakšgadījums — nsubj:pass (passive nominal subject) — pasīva subjektiem. **Nepieciešams izanalizēt, vai mums ir morfoloģisks pamats šķirt?**

4.2.28. *nummod*: numeric modifier

—

4.2.29. *obj*: direct object

—

4.2.30. *obl*: oblique nominal

Nomināli paplašinātāji, kas modificē predikātus. Šajā grupā ietilpst situanti un determinanti. UDv1 bija daļa no *nmod*. **Vai uz mums attiecas *obl:agent*?**

4.2.31. *orphan*: orphan

Šobrīd tiek likts gadījumos, ja pie reducēta izteicēja nav *aux* vai *cop* un reducēts kaut kas tāds, kas UD iekristu sadaļā *Core arguments*.

4.2.32. *parataxis*: parataxis

Vai šīs varētu būt neatkarīgās teikuma daļas?

4.2.33. *punct*: punctuation

—

4.2.34. *reparandum*: overridden disfluency

Runātāja paša izteikts labojums vai divas reizes atkārtots viens un tas pats vārds slikti rediģētā tekstā. LVTB šobrīd ir tikai otrais no minētajiem gadījumiem.

4.2.35. *root*: root

Izteicēja vai izteikuma pamatvārda atkarība no mākslīgas koka virsvirsotnes. UD vadlīnijas pieprasa, lai kokā būtu tikai viena virsotne ar šādu lomu. Ja izteicējs (vai izteikuma pamatelemements) ir izlaists (redukcija) un tā vietā ir vairāki "bārenīši", tad UD vadlīnijas liek to, kurš atrodas vistālāk pa kreisi no visiem, mākslīgi padarīt par vienīgo *root*.

4.2.36. *vocative*: vocative

Uzruna — šķiet, ka pilnībā atbilst uzrunai LVTB izpratnē.

Tomēr jāpārliecinās, vai ar uzrunām "tu, Jāni, labāk…" viss ir kārtībā.

4.2.37. *xcomp*: open clausal complement

Klauzāls verba vai adjektīva atkarīgais, kam nevar būt pašam savs subjekts — subjektu neizbēgami nosaka tas, kas ir kokā virs šī *xcomp*. Šo lomu lieto tikai pēc valences obligātajiem paplašinātājiem. **Nepieciešams pieskatīt, lai mēs nesamarķējam neobligātās lomas šādi. Vai, ja to nevaram, tad vismaz godīgi atrunāt.**

Šo lomu nelieto lietvārda atkarīgajiem (tiem ir *acl*?)

4.3. Redukcijas

Citāts no UD vadlīnijām https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis:

> *The UD approach to ellipsis can be summarized as follows:*
>
> 1. *If the elided element has no overt dependents, we do nothing.*
> 2. *If the elided element has overt dependents, we promote one of these to take the role of the head.*
> 3. *If the elided element is a predicate and the promoted element is one of its arguments or adjuncts, we use the orphan relation when attaching other non-functional dependents to the promoted head.*

Redukcijas virsotnes, kam atbilst teikumā esoša tekstvienība, sadala divās virsotnēs: (1) tukšā redukcijas virsotnē ar redukcijas lauka informāciju un līdzšinējo lomu un (2) virsotnē, kas satur tekstvienību, tās morfoloģisko informāciju, kārtas numuru teikumā un lomu *ellipsisTok*. Ja sadalāmā virsotne ir xVārda vai PMC daļa, tad abas sadalītās virsotnes kļūst par attiecīgā xVārda vai PMC daļām. Ja sadalāmā virsotne ir atkarīgais vai koordinācijas konstrukcijas daļa, tad *ellipsisTok* kļūst par redukcijas virsotnes atkarīgo. Tālāk pārveidošana notiek atbilstoši *UD mapping.odg* un iepriekšējā nodaļā atrunātajam. Mākslīgi radītā loma *ellipsisTok* kļūst par *punct*, ja morfotags ir z.*. **Par ko tā**

kļūst citos gadījumos? Lai transformācija strādātu korekti, vajag, lai gadījumos, kad *ellipsisTok* kļūst par atkarīgo, virsotnei būtu arī citi atkarīgie.

Redukcijas virsotnes, kam teikumā neviena tekstvienība neatbilst un nav atkarīgo, piemēram, reducēta saitiņa saliktā izteicējā, šobrīd ignorē pilnībā.

Redukcijas virsotnes, kam teikumā neviena tekstvienība neatbilst, bet atkarīgie ir, pārveido, izvēloties vienu no atkarīgajiem un "ieceļot" par pārējo atkarīgo vecāku.

Ja izlaists nomens: *amod* > *nummod* > *det* > *nmod* > *case*. No savas pieredzes vēl galā (ar pēdējo prioritāti) pielikām *acl* "iespēju noskatīties, otrkārt [iespēju] izveidot". **Pašlaik atribūta gadījumā ņem pirmo. Vai var UD diskusijās atrast kaut ko, kas ļauj tā nedarīt?**

Ja izlaists izteicējs (*pred*) vai darbības vārds finītā formā (v..[^pn].*) un ir pieejams *aux*, *cop* vai *mark* nenoteiksmes gadījumā (mums nav), tad lieto kādu no tiem. **Kādā secībā, ja ir gan *aux*, gan *cop*? Pašlaik ņem to, kas gadās pirmais.** Ja palīgdarbības vārdu nav vai ja izlaists *spc*, kas izteikts ar v..pu.*, v..pp.* vai v..n.*, paceļ prioritāšu secībā kādu no *nsubj* > *obj* > *iobj* > *obl* > *advmod* > *csubj* > *xcomp* > *ccomp* > *advcl* > *dislocated* > *vocative*. Ja izlaista *pred*, finītas formas verba (v..[^pn].*) vai ar v..pu.*, v..pp.* vai v..n.* izteikta *spc* vietā paceļ kādu no šiem, tad izlaistā elementa nepaceltie atkarīgie, kas sanāk UD nefunkcionālajām lomām, iegūst lomu *orphan*. To, kas var iegūt lomu *orphan*, sašaurināja UD diskusija [UD docs #643]: *nsubj*, *nsubj:pass*, *obj*, *iobj*, *csubj*, *csubj:pass*, *ccomp*, *xcomp*, *obl*, *vocative*, *dislocated*, *advcl*, *advmod*.

Ja reducētajai virsotnei ir tieši viens bērns, tad to paceļ uz augšu, pat tad ja reducētās virsotnes loma nav aprakstīta.

Ja ļoti reducētā teikumā ir palicis tikai determinats vai situants un saiklis, tad patiesībā laikam labāk būtu, ja redukcijas vietu aizpildītu determinats/situants, nevis saiklis, kā šobrīd sanāk (tāpēc, ka teikuma daļas kopatkarīgie netiek izskatīti reducētā izteicēja vietas aizpildīšanai)

4.4. Paplašinātās atkarības

4.4.1. Redukcija

Redukcijas virsotnēm, kam ir bērni, tiek izveidotas tukšas atkarību koka virsotnes. **UD prasa šādi darīt tikai predikātiem, kā labāk identificēt? Pašlaik pārbauda, vai redukcijas tags ir v..[^p].* vai v..p[ud].*.** Tagu tām nosaka pēc redukcijas laukā ievietotā taga. Formu ņem no redukcijas lauka iekavām. Lemmu paprasa morfoservisam pēc taga un formas. Var gadīties, ka kāda lauka trūkst.

4.4.2. Vienlīdzīgi teikuma locekļi

Katram vienlīdzīgajam teikuma loceklim pievieno šādas saites:

1. uz vecāku atkarību kokā,
2. no koordinācijas kopīgajiem atkarīgajiem,

3. no frāžu daļām, kas atkarību kokā sanāk atkarīgas no šī frāzes elementa, kas izteikts ar koordināciju.

   Netiek vilktas saites uz/no PMC elementiem ar lomu *punct*.

4.4.3. Kontrolētie un paaugstinātie teikuma priekšmeti

Salikta izteicēja daļām piesaista izteicēja kopīgo teikuma priekšmetu *subj* vai teikuma priekšmeta palīgteikumu *subjCl*. Papildu teikuma priekšmetus nekad nepiesaista lemmām *būt*, *tikt*, *tapt*~~*, kļūt*~~ *xPred auxVerb* lomā. Teikuma priekšmetu piesaistīšanu veic tikai teikuma priekšmetiem, kam UD paplašināto atkarību loma pret vecāku sanāk *nsubj*, *nsubj:pass*, *csubj*, *csubj:pass*.

Nomināliem subjektiem papildsubjekta lomu izvēlas pēc verba vai xPred taga (izvēlas lielāko vēl attiecināmo vienību): v..[^p].....p.*, v[^\[]*\[pas.* un v..pd...p.* nosaka *nsubj:pass* vai *csubj:pass*, pārējie verbu tagi (v.*) nosaka *nsubj* vai *csubj*. Izvēle starp nominālu un klauzālu subjektu tiek veikta, vadoties pēc teikuma priekšmeta sākotnējās lomas — ja sākotnējā loma sanāk *nsubj* vai *nsubj:pass*, tad arī papildsaitēm izvēlas starp lomām *nsubj* un *nsubj:pass*, savukārt, ja *csubj* vai *csubj:pass* — tad starp *csubj* un *csubj:pass*.

Ko darīt, ja vajag piesaistīt ne nominālu subjektu, piemēram, infinitīvu? Ko darīt, ja jāpiesaista izteicēja daļai, kas nav verbs?

TODO: sataisīt papildsubjektu saites starp -dams divdabjiem bez sava teikuma priekšmeta un attiecīgajiem teikuma priekšmetiem.

4.4.4. Locījumi

UD lomām *nmod*, *obl*, *acl* un *advcl* pievieno locījumu, ja izteikts ar nomenu, vai prievārdu/saikli, ja izteikts ar *xPrep*/*xSimile*. Ja *xSimile* saiklis ir *xFunctor*, ņem secīgi pēdējā *basElem* lemmu.

TODO: ko darīt ar palīgteikumiem un ar (daļēji) lokāmu divdabi vai nenoteiksmi izteiktiem *nmod*, *obl*, *acl* un *advcl*?

4.4.5. Palīgteikumu references

TODO: kā atšķirt palīgteikumu ievadošos vārdus no visa cita?

5. Pēcapstrāde

Pēc visa koka pārveidošanas tiek veikta pēcapstrāde pieturzīmju novietojuma precizēšanai:

1. Kamēr kokā atrodamas vietas, kur *punct* pakārtots *punct*, zemākais no tiem tiek pārcelts vienu līmeni augstāk. Šāda pēcapstrāde nepieciešama tādiem iespraudumiem kā *[..]* un *(!)*.
2. **Iespējams, ka mums vajadzēs vēl kaut ko neprojektīvām tiešo runu pieturzīmēm.**

Avoti

[Nivre, et al, LREC] Nivre, de Marneffe, Ginter, Goldberg, Hajič, Manning, McDonald, Petrov, Pyysalo, Silveira, Tsarfaty, Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. Iesniegts LREC 20016, http://stp.lingfil.uu.se/pipermail/ud/attachments/20151026/37b100f8/attachment.pdf

[ud-web-morfo] https://universaldependencies.github.io/docs/u/overview/morphology.html

[UD docs #643] https://github.com/UniversalDependencies/docs/issues/643 — diskusija par *orphan* lomām un redukcijas pacelšanu no tehniskā viedokļa

[UD typos] https://universaldependencies.org/u/overview/typos.html — apraksts par kļūdu labošanu