Manne dihtor galgá máhttit grammatihka?

LENE ANTONSEN – TROND TROSTERUD

Ođđaáigásaš servodagas ii oktage giella birge čállingiellan giellateknologiija haga. Sámegillii leat ráhkaduvvon grammatihkalaš analysáhtorat, vaikko erenoamážit stuora gielaide lea dábálaš geavahit statistihkalaš analysáhtoriid. Artihkal gieđahallá man muttus sámi giellateknologiija lea dál, mo sámegiella earrána dárogielas ja eŋgelasgielas, ja mo dat erohusat váikkuhit giellateknologalaš čovdosiidda.

Vuosttaš kapihttalis čilgejetne lokaliserema, dahje vuođđoinfrastruktuvrra mii dahká vejolažžan čállit sámegillii dihtoriin. Dasto guorahalle mo sátneanalysas golbma iešguđetlágan lahkonanvuogi heivejit sámegillii (kap. 2.). Goalmmát kapihttalis gehčče grammatihkalaš cealkkaanalysa, ja de čuovvovaš kapihttalis makkár geavatlaš prográmmaide sáhttá analysáhtoriid atnit odne juo. Maŋimuš fáddá (kap. 5) lea mo giellateknologiija váikkuha sámegiela geavaheapmái min servodagas. Guovddážin lea ahte dakkár reaiddut šaddet dehálažžan jus galgá leat vejolaš geavahit giela omd. hálddahuslaš giellan.

1. Lokaliseren

Dihtormáilmmis tearbma *lokaliseren* máksá buot maid dihtoris ferte heivehit dihto riikii ja gillii. Sámegiela lokaliseremii gullá áigediehtu, alfabehtalaš sorteren, boallobeavdi ja erenoamáš bustávat.

Buot sámegielaid čállingielain gávdnojit bustávat mat eai gávdno riikkaid váldogielain, nugo boahtá ovdan tabeallas 1.

Tabealla 1. Sámegielaid erenoamáš bustávat.

Giella	Alfabehta	Liigebustávat
Lullisámegiella	latiinnalaš	ï
Julevsámegiella	latiinnalaš	á, ŋ (= ń)
Davvisámegiella	latiinnalaš	á, č, đ, ŋ, š, ŧ, ž
Anársámegiella	latiinnalaš	â, á, ä, č, đ, ŋ, š, ž
Nuortalašgiella	latiinnalaš	â, ä, å, č, ʒ, ǯ, đ, ǧ, ɡ, ǩ, ŋ, õ, š, ž, ´
Gielddasámegiella	kyrillalaš	ā, ӓ, ē, ӣ, ӣ, ԯ, ӎ, ӈ, ӊ, ō, ҏ, ӯ, ӛ, ә̄, ю̄, я̄, ҍ

Sámegiela bustávat leat guktii integrerejuvvon dihtormáilbmái: Vuosttaš geardi lei ovdal 1995. Juohke geavaheaddjis lei sierra dihtorčoavddus, vai sáhtii čállit, lohkat ja čálihit báhpárii dihtoriin. Dalle ledje máŋga iešguđetlágan dihtorčovdosa. Seammá bustáva duohken lei juohke čovdosis iežas nummirkoda. Nu guhká go ulbmil lei čálihit teavsttaid báhpárii, de variašuvdna ii dagahan váttisvuođaid.

Jagi 1995 maŋŋel, interneahttaáiggis, lei dárbu gávdnat oktasaš čovdosa. Ulbmil lei dál sáddet šleađgaboastta, sirdit sámegiel fiillaid dihtoris nubbái ja interneahtas lohkat teavsttaid. Sullii jagi 2000 rájes lea riikkaidgaskasaš dihtorstandárda Unicode leamaš dábálaš interneahtas. Unicode-standárddas gávdnojit prinsihpas buot máilmmi gielaid bustávat. Sámegiela bustávat leat leamaš mielde Unicode-standárddas jagi 1993 rájes.

Okta sámi bustávain, nammalassii *đ*, lea váttis vel dán áigge. Váttisvuohta ii leat teknihkalaš, muhto lea čadnon geavaheddjiid máhttui. Jos čálli ii geavat sámi boallobeavddi, de son dávjá geavaha islándalaš bustáva *ð*. NRK (*Norsk rikskringkasting*) podcast-listtus seammá prográmmanamma molsašuddá beaivvis beaivái: *Ođđasat, Oððasat, Oddasat*. Interneahtas seammá fenomena gávdno dieđalaš máilmmis. Google-ohcanprográmmain gávdná golbma sierra tearpma: *Dieđut, Dieðut, Diedut* (gč. tabealla 2). Sátni *dieđut* lea dieđusge maiddái dábálaš sátni sámegielas.

Tabealla 2. Mo sánit Dieđut/dieđut ja Ođđasat/ođđasat leat čállojuvvon interneahtas 20.11.09.

Ohcantearbma	***dieđut***	***dieđut***	***dieđut***	***ođđasat***
Ohcandomena	**samiskhs.no**	**uit.no**	**.no**	**olles interneahtta**
Dárogiela d	0,7 %	30,4 %	3,9 %	9,0 %
Islánddagiela ð	56,5 %	22,2 %	2,7 %	86,0 %
Sámegiela đ	42,8 %	47,4 %	93,4 %	5,0 %

Go rivttes bustávva lea geavahuvvon *dieđut*-sátnái *.no*-domenas 93,4 % dáhpáhusain, muhto *ođđasat*-sátnái dušše 5,0 %, de lea sivva ahte *dieđut* lea geavahuvvon sámegiel teavsttain, ja *Ođđasat* lea siterejuvvon TV-prográmma-namman (boasttu bustávain) dárogiel teavsttain.

Nubbi lokaliserenváttisvuohta ii leat diehtemeahttunvuođas, muhto ieš dihtorvuogádagas. Sullii jagi 2000 maŋŋel buot dihtorvuogádagat dorjot sámegiela, muhto servodaga stuora registarat Norggas, nugo Álbmotregisttar ja Brønnøysund-registarat[1], geavahit velge boares dihtorstandárddaid, sámegiela bustávaid haga. Álbmotregistaris livčči dehálaš geavahit á (*Ánde, Márjá*, jna.), ja dat ii leat mihkkege čuolmmaid teknihkalačča t, muhto Álbmotregisttaris lea mearriduvvon geavahit dušše dáru, ruoŧa ja duiskka bustávaid. Brønnøysund-registaris geavahuvvo ng. *Latin 1* -kodatabealla, sámegiela bustávaid haga (earret *á*). Ádjána vel máŋga jagi ovdal go Brønnøysund-registariidda lea vejolaš registreret fitnodatnama nugo *Šillju Bistro*.

Lokaliseremii gullá maiddái vejolašvuohta čujuhit iešguđet gillii. Dál juohke sámegielas lea standárda ISO-koda[2] (lulli-, julev- ja davvisámegiela ISO-639-kodat leat *sma, smj, sme*). Microsoft-fitnodat ii geavat ISO-kodaid, muhto baicca iežas giellakodaid. Windows MS-Office-páhkas lea vejolaš čujuhit buot sámegielaide, muhto Macintosh MS-Office-páhkas eai gávdno giellakodat. Dan dihte lulli-, julev- ja davvisámegiel divvunprográmmaid ferte ohcat slováhka-, euskara- (baska-) ja katalánagiela vuolde.

---

1 Našunála dárkkistan- ja registrerenvuogádat, mii erenoamážit guoská fitnodagaide ja servviide.

2 ISO = International Organization for Standardization

2. Sániid analyseren

Sátni, mii lea muhtun teavsttas oassin, ii leat dušše sátni, muhto sátni ovddasta dihto leksema, grammatihkalaš sáni ja gullá dihto sojahanparadigmii. Gávdnojit máŋggalágan giellateknologiijat sátneanalysa várás – juohke teknologiija heive dihto geavahussii ja dihto gielaide. Dás guorahalle máddagastima, statistihkalaš lahkonanvuogi ja morfologalaš analysáhtora.

2.1. Máddagastin

Lunddolaš gielain seammá leksemas sáhttet leat máŋga sojahanhámi. Go ohcá dihto leksema teavsttas dahje dokumeanttain, de galgá leat vejolaš gávdnat maiddái sojahuvvon hámiid. Sátnehámit *gahpira*, *gahpiris*, *gahpiriin* leat buot čadnon *gahpir*-leksemii, seamma ládje go dárogiel sánit *hatt, hatten, hattar, hattane, hattens, hattanes* gullet oktii.

Dábálaš vuohki lea máddagastin (eŋg. *stemming* ʻmáddaga dahkatʼ). Dárogillii máddagastinprográmma válddášii eret gehčosiid *-en, -ar, -ane, -ens, -anes*, ja dalle bázášii *hatt* máttan. Máddagastin doaibmá maiddái bures agglutinerejeaddji gielas, dego omd. turkkagielas, mas olles paradigmas lea seammá mátta (*şapka* ʻgahpirʼ: *şapka-nınm şapka-ya, şapkay-ı, şapka-da, şapka-dan, ...*).

Sámegillii lea maiddái vejolaš ráhkadit máddagasti, muhto sámegiela morfologiija lea earálágan go turkkagiela morfologiija go paradigmas sáhttet dássemolsašumi ja diftoŋganjuolgama dihte leat máŋga máddaga. Dán artihkkala oktavuođas moai letne ráhkadan sámegiela máddagasti, ja dainna analyseren substantiivvaid ja vearbbaid sojahanparadigmaid[3]. Substantiivvain leat máddagastima maŋŋel gaskamearálaččat 3,1 máddaga, ja vearbbain 4,9 máddaga.

Gávdnan dihte man stuora čuovvumušat das leat teakstaanalysii, de letne iskan analyseret vuosttaš 100.000 sáni Norgga Almmolaš Čielggadeami 1994–19:s (NAČ). Morfologalaš analysain (gč. Kap. 2.3) teavsttas leat

[3] Substantiivvat leat *baste, beana, boazu, fális, gahpir, geažus, gieddi, mánná, nieida, reŋko, sabet, viessu*, ja vearbbat leat *viežžat, diehtit, čierrut, boradit, čohkkát, gillet, diŋgot*. Máddagasti lea interneahtas: https://victorio.uit.no/langtech/trunk/gt/sme/src/sme-stemmer.xfst

5.128 substantiivaleksema. Máddagasti mielde seammá teavsttas leat 9.108 substantiivamáddaga. Dat mearkkaša ahte 44 % máddagiin eai čadnojuvvo rivttes leksemii.

Máddagastinproseassa lea vejolaš buoridit, muhto sámegielat leat dasa hástalussan. Sámegielain leat rikkis paradigmat, maid lea bággu gieđahallat jus galgá gávdnat rivttes leksema. Sámegielain leat siskkáldas morfofonologalaš proseassat, dássemolsašupmi, vokálarievdan ja diftoŋganjuolgan, maid dihte buhtes morfologalaš máddagastima maŋŋel juohke sánis leat máŋga máddaga. Jos ulbmil lea juohke leksemii gávdnat buot sojahanhámiid, de máddagastin ii heive sámegillii.

2.2. Statistihkalaš analysáhtor

Máŋga giellateknologalaš prográmma eŋgelasgiela várás leat ráhkaduvvon unnán grammatihkalaš analysaiguin. Čállindárkkistanprográmmaide leat čohkken buot iešguđetlágan sátnehámiid stuora dárkkistuvvon teakstačoakkáldagas, analysa haga, ja dien ládje prográmma dovdá «buot» normatiiva čállinhámiid eŋgelasgillii.

Guokte beali leat dahkan dán vuogi dohkálaš vuohkin: Vuosttažettiin leat ollu eŋgelasgiel teavsttat digitálalaš hámis viežžan ládje. Dasa lassin lea eŋgelasgielas geafes morfologiija. Juohke leksemii gullet dušše moadde sátnehámi, iige oktage dáin leat ollu hárvvit go dat eará hámit. Eanaš gielain máilmmis lea mohkkát struktuvra go eŋgelasgielas – struktuvra mas leksemii gullet viiddis sojahanparadigmat máŋggainlogiin dahje máŋggainčuđiin sátnehámiin, ja muhtun sátnehámit sáhttet leat hárvenačča t. Statistihkalaš lahkonanvuohki doaibmá buoremusat gielain main gávdnojit hui ollu digitálalaš teavsttat ja unnán morfologiija. Čájehan dihte manne lea nu, de letne iskan guovtti vearbba sojahanhámiid frekveanssa eŋgelasgiel, dárogiel ja sámegiel ođđa testameanttas. Nubbi vearba lea hui frekveanta, nubbi lea gaskamearálaš frekveanta (gč. tabealla 3).

Tabealla 3. Ođđa testameantta sturrodat golmma gillii – ja vearbbat maid frekveanssa letne iskan.

Ođđa testameanta	**galle sáni**	**hui frekveanta vearba**	**gaskamearálaš frekveanta vearba**
davvisámegillii (OT)	138.706	*dadjat*	*bálvalit*
dárogillii (DNT)	137.148	*si*	*tjene*
eŋgelasgillii (NT)	188.616	*say*	*serve*

Tabealla 4 čájeha gallii dárogiel ja eŋgelasgiel vearbbaid sátnehámit gávdnojit ođđa testameanttas. Passiivahámit eai leat mielde danne go sámegielas passiivavearbbat gullet suorggideapmái eaige sojaheapmái, nugo eŋgelasgielas ja dárogielas. Homonymat leat tabeallas mielde dušše oktii, nuppi ruktui lea merkejuvvon x .

Tabealla 4. Guovtti vearbba frekveansa dárogiel ja eŋgelasgiel ođđa testameanttas. x = homonyma eará sátnehámiin.

Dárogiel ođđa testameanttas (DNT)					**Eŋgelasgiel ođđa testameanttas (NT)**				
Preseansa	*sier*	467	*tjener*	26	**Preseansa**	*say*	422	*serve*	33
					Sg2	*sayest*	39	*servest*	2
					Sg3			*serveth*	9
Preterihtta	*sa*	1028	*tjente*	7	**Preterihtta**	*said*	1058	*served*	5
Imperatiiva	*si*	199	*tjen*	2	**Imperatiiva**	*say*	x	*serve*	x
Infinihtta hámit					**Infinihtta hámit**				
Perfeakta partisihppa	*sagt*	136	*tjent*	11	**Perfeakta partisihppa**	*said*	x	*served*	x
Infinitiiva	*si*	x	*tjene*	24	**Infinitiiva**	say	x	*serve*	x
					Gerunda	*saying*	408	*serving*	6

Eŋgelasgiela s-hámit *serves* ja *says* eai gávdno danne go dát veršuvdna Ođđa testameanttas (NT) lea čállojuvvon jagis 1611, ja das leat 2. ja 3. persovdna *-st* ja *-th,* eaige leat 3. persovdna *s*-hámit. Go ohcá dábálaš dán áigásaš teakstačoakkáldagas mas leat 85.000 sáni (bealli ođđa testameanttas), de *says* lea 4 gearddi ja *serves* lea 2 gearddi. Boađus lea ahte buot dán guovtti

vearbba sojahanhámit gávdnojit eŋgelasgiel ođđa testameantta sturrosaš teakstačoakkáldagas. Seammá guoská dárogillii, go dáru ođđa testameanttas leat buot sojahanhámit.

Miellagiddevaš lea buohtastahttit dán bohtosa sámegielain. Eŋgelasgielas ja dárogielas eai leat go 2–3 sierralágan finihtta hámi, dasa lassin bohtet 1–2 infinihtta hámi. Davvisámegielas leat juohke vearbbas 42 sierralágan finihtta hámi ja dasa lassin 7 infinihtta hámi. Dás eat geahča suorggádusaid go juohke suorggádussii gulašii fas sojahanparadigma. Tabeallas 5 leat *dadjat*-vearbba sátnehámit mat gávdnojit ođđa testameanttas (OT).

Tabealla 5. dadjat-vearbba frekveansa davvisámegiel ođđa testameanttas (OT). Go sátnehápmi ii gávdno, de ruktu lea suivejuvvon. x = homonyma eará sátnehámiin.

	Indikatiiva preseansa		**Indikatiiva preterihtta**		**Konditionála**		**Potentiála**		**Imperatiiva**	
Sg1	*dajan*	4	*dadjen*	0	*dajašin*	1	*dajažan*	0	*dadjon*	0
Sg2	*dajat*	5	*dadjet*	x	*dajašit*	0	*dajažat*	0	*daja*	x
Sg3	*dadjá*	42	*dajai*	204	*dajašii*	4	*dajaža*	0	*dadjos*	0
Du1	*dadje*	182	*dajaime*	0	*dajašeimme*	0	*dajažetne*	0	*daddju*	0
Du2	*dadjabeahtti*	1	*dajaide*	0	*dajašeidde*	0	*dajažeahppi*	0	*daddji*	4
Du3	*dadjaba*	0	*dajaiga*	7	*dajašeigga*	0	*dajažeaba*	0	*dadjoska*	0
Pl1	*dadjat*	45	*dajaimet*	1	*dajašeimmet*	0	*dajažit*	0	*dadjot*	0
Pl2	*dadjabehtet*	24	*dajaidet*	0	*dajašeiddet*	2	*dajažehpet*	0	*daddjet*	0
Pl3	*dadjet*	47	*dadje*	182	*dajašedje*	0	*dajažit*	x	*dadjoset*	0
Neg	*daja*	3			*dajaše*	1	*dajaš*	1	*daja*	x
	Infinitiiva		**Aktio essiiva**		**Perf.part.**		**Vearba-abessiiva**		**Gerunda**	
	dadjat	x	*dadjamin*	1	*dadjan*	14	*dajakeahttá*	0	*dajadettiin*	0

Muhtun grammatihkalaš sániin leat guokte dahje eanet vejolaš sátnehámi. Tabeallain lea dalle mielde sátnehápmi mii gávdno guorahallojuvvon teavsttas.

Nugo boahtá ovdan tabeallas 5, de ođđa testameanttas gávdná dušše 22 sierralágan sojahanhámi *dadjat*-vearbbas, ja dat dahká 44 % buot hámiin. Buoremusat gokčá indikatiivva preseanssa, muhto dátge vuolleparadigma báhcá váilevažžan.

Tabealla 6 čájeha davvisámegiel *bálvalit*-vearbba bohtosiid. Ođđa testameanttas leat dušše 15 sierralágan sojahanhámi, mat dahket 31 % vejolaš sátnehámiin.

Tabealla 6. bálvalit-vearbba frekveansa davvisámegiel ođđa testameanttas (OT). Go sátnehápmi ii gávdno, de ruktu lea suivejuvvon. x = homonyma eará sátnehámiin.

	Indikatiiva preseansa		**Indikatiiva preterihtta**		**Konditionála**		**Potentiála**		**Imperatiiva**	
Sg1	*bálvalan*	13	*bálvalin*	0	*bálvalivčče**n***	0	*bálvaleačča**n***	0	*bálvalehko**n***	0
Sg2	*bálvalat*	0	*bálvalit*	x	*bálvalivčče**t***	0	*bálvaleačča**t***	0	*bálval*	x
Sg3	*bálvala*	10	*bálvalii*	3	*bálvalivčči**i***	0	*bálvaleažžá*	0	*bálvalehkos*	2
Du1	*bálvaletne*	0	*bálvaleimme*	1	*bálvalivčči**ime***	0	*bálvaleže*	0	*bálvaleadnu*	0
Du2	*bálvaleahppi*	0	*bálvaleidde*	0	*bálvalivčči**ide***	0	*bálvaleažžabeahtti*	0	*bálvaleahkki*	0
Du3	*bálvaleaba*	0	*bálvaleigga*	0	*bálvalivčči**iga***	0	*bálvaleažžaba*	0	*bálvalehkoska*	0
Pl1	*bálvalit*	0	*bálvaleimmet*	0	*bálvalivčči**imet***	0	*bálvaleažžat*	0	*bálvalehkot*	1
Pl2	*bálvalehpet*	0	*bálvaleiddet*	1	*bálvalivčči**idet***	0	*bálvaleažžabehtet*	0	*bálvalehket*	2
Pl3	*bálvalit*	x	*bálvaledje*	3	*bálvalivčče*	0	*bálvaležžet*	0	*bálvalehkoset*	1
Neg	*bálval*	5			*bálvalivčče*	x	*bálvaleaš*	0	*bálval*	x
	Infinitiiva		**Aktioessiiva**		**Perf.part.**		**Vearbaabessiiva**		**Gerunda**	
	bálvalit	38	*bálvaleamen*	2	*bálvalan*	x	*bálvalkeahttá*	0	*bálvalettiin*	1

Sámegiel ođđa testameanttas leat 139.000 sáni, juoga mii lea unnán dán oktavuođas. Interneahtta dahká stuorit teakstačoakkáldaga, dál leat sullii 75 milj. sámegiel sáni interneahtas[4] (skábmamánus 2009). Tabeallas 7 leat merkejuvvon sátnehámit mat leat interneahtas. Dan dihte go muhtun sátnehámit leat homonymat eará gielaid sátnehámiiguin, de ii leat frekveansa mielde tabeallas. 48 sierralágan hámis gávdnojedje 20 interneahtas, ja dat dahká 42 %. Giellatekno (Romssa universitehta sámi giellateknologiija guovddáš) ja Norgga Sámedikki Divvun-joavkku fiillat mat leat olámuttos interneahta bokte, eai leat mielde tabeallas, danne go muhtun fiillain leat genererejuvvon aiddo dáid vearbbaid sátnehámit.

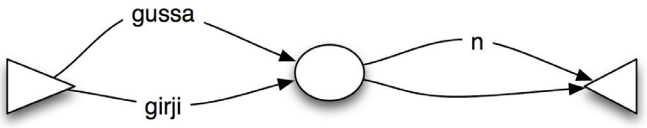
[4] Interneahta sámegiela teakstamearri lea árvvoštallojuvvon ná: Giellatekno/Divvuma davvisámegiela teakstačoakkáldagas (6,997 milj. sáni) rehkenasttiime man stuora proseantaoasi 34 dábálaš sátnehámi dahke (*bierggu, boahtibehtet, boahtán, bohtet, ...*). Bijaime vuođđun ahte sániid proseantaoassi lea seammá interneahtas, ja nu árvvoštalaime interneahta teakstameari.

Tabealla 7. bálvalit-vearba interneahtas (01.12.09). Go sátnehápmi ii gávdno, de ruktu lea suivejuvvon. +: sátnehápmi gávdno, 0: ii gávdno, x = homonyma eará sátnehámiin.

	Indikatiiva Preseansa		**Indikatiiva Preterihtta**		**Konditionála**		**Potentiála**		**Imperatiiva**	
Sg1	*bálvalan*	+	*bálvalin*	0	*bálvalivččen*	+	*bálvaleačča*n	0	*bálvalehkon*	0
Sg2	*bálvalat*	+	*bálvalit*	x	*bálvalivčče*t	0	*bálvaleačča*t	0	*bálval*	x
Sg3	*bálvala*	+	*bálvalii*	+	*bálvalivčči*i	+	*bálvaleažžá*	0	*bálvalehkos*	+
Du1	*bálvaletne*	0	*bálvaleimme*	0	*bálvalivččiime*	0	*bálvaležže*	0	*bálvaleadnu*	0
Du2	*bálvaleahppi*	0	*bálvaleidde*	0	*bálvalivččiide*	0	*bálvaleažžabeahtti*	0	*bálvaleahkki*	+
Du3	*bálvaleaba*	+	*bálvaleigga*	0	*bálvalivččiiga*	0	*bálvaleažžaba*	0	*bálvalehkoska*	0
Pl1	*bálvalit*	+	*bálvaleimmet*	0	*bálvalivččiimet*	0	*bálvaleažžat*	0	*bálvalehkot*	+
Pl2	*bálvalehpet*	+	*bálvaleiddet*	0	*bálvalivččiidet*	0	*bálvaleažžabehtet*	0	*bálvalehket*	+
Pl3	*bálvalit*	x	*bálvaledje*	+	*bálvalivčče*	+	*bálvaležžet*	0	*bálvalehkoset*	0
Neg	*bálval*	+			*bálvalivčče*	x	*bálvaleaš*	0	*bálval*	x
	Infinitiiva		**Aktio essiiva**		**Perf.part.**		**Vearbaabessiiva**		**Gerunda**	
	bálvalit	x	*bálvaleamen*	+	*bálvalan*	x	*bálvalkeahttá*	+	*bálvalettiin*	+

2.3. Morfologalaš analysáhtor

Gielaide main leat unnán digitálalaš teavsttat ja ollu morfologiija, lea čoavddus ráhkadit automáhtaid mat lasihit gehčosiid máddagiidda. Govvosis 1 automáhta addá sátnehámiid *gussa*, *gussan*, *girji*, *girjin*.

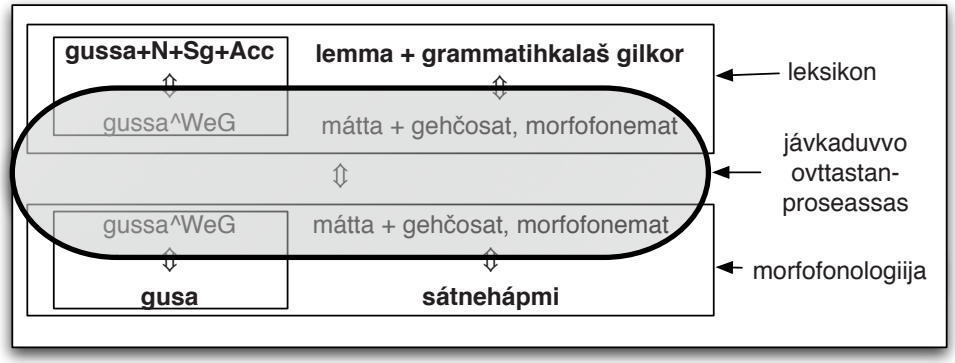


Govus 1. Morfologalaš automáhta.

Grammatihkalaš transduser (Beesley – Karttunen 2003) lea dihto automáhtatiipa mas juohke sátnehámis leat guokte ovddasteami. Sátnehámis *gussan* lea maiddái ovddasteapmi «gussa+N+Ess». Sátnehápmi lea atta (eŋg. *input*) ja leksema + grammatihkalaš gilkorat (= grammatihkalaš sátni) fas buvttus (eŋg. *output*), dahje nuppe ládje. Dákkár transduser gohčoduvvo loahpalaš dilletransduserin – FST (eŋg. *finite state transducer*).

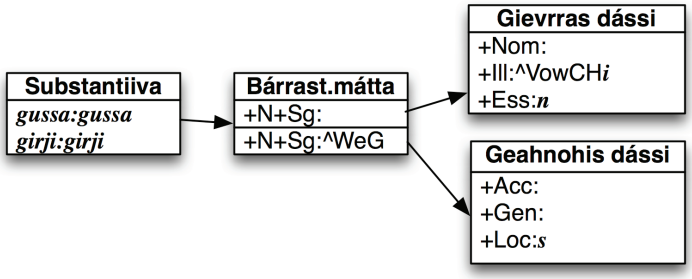
Turkkagielas lokatiivatransduser sátnái *kitap* ‘girji’ sáhttá leat transduserat [kitap:kitap] ja [+Sg+Loc:da], nu ahte atta lea *kitap+Sg+Loc*, ja buvttus *kitapda*. Nugo oinniimet kapihttalis 2.1. máddagastima birra, de dakkár transduser doaimmašii bures turkkagillii. Dat ii doaimmaše bures sámegillii, go jus atta lea *girji+Sg+Loc*, de buvttus livčči *girjis*.

Dan dihte sámegiela analysáhtoris leat guokte transdusera: Nubbi leksikonii ja geažusmorfologiijai (leksikontransduser), ja nubbi siskkáldas morfofonologalaš proseassaide (morfofonologalaš transduser, gč. Moshagen – Sammallahti – Trosterud 2004). Leksikontransdusera bajit dássi ovddasta grammatihkalaš sáni, ja vuolit dásis leat mátta, affivssat ja báhcahagat morfofonologalaš proseassaide maid boađus lea sátnehápmi. Govus 2 čájeha leksikon- ja morfofonologalaš transduseriid. Leksikontransdusera buvttus lea morfofonologalaš transdusera atta.



Govus 2. Analysáhtoris leat guokte transdusera, nuppis lea leksikon ja gehčosiid lasiheapmi, nuppis leat morfofonologalaš proseassat.

Govvosis 2 sátnehápmi *gusa* lea ovdamearkan. Go leksikon- ja morfofonologalaš transduserat leat ovttastuvvon, de gaskabuvttus jávká. Boađus lea transduser, man vuolit dássi lea *gusa* ja bajit dássi lea *gussa+N+Sg+Acc.* Transduseris bajit ja vuolit dásit sirrejuvvojit duppalčuoggáin (gusa:gussa+N+Sg+Acc), muhto loahpalaš buvttus lea dan haga, nugo govvosis 5.



Govus 3. Leksikontransduser. Báhcahagat leat ^WeG ⟶ geahnohis dássi ja ^VowCH ⟶ vokála rievdan. Bajit dási segmeanttat leat duppalčuoggá gurutbealde, ja vuolit dási fas duppalčuoggá olgešbealde.

Leksikontransdusera siskkáldas struktuvra lea govvosis 3, mii čájeha ahte *gussa-* ja *girji*-leksemaid bajit dássái lasihuvvojit sátneluohkká- ja kásusgilkorat. Vuolit dássái lasihuvvojit báhcahagat (eŋg. *trigger*) dássemolsašuddama ja soggevokála rievdama várás, ja dasa lassin geažus.

Dássemolsašuddan ja vokálarievdan dahkko juogo fonologalaš konteavstta vuođul dahje lasihuvvon báhcahaga dihte. Govvosis 3 lea WeG geahnohis dási báhcahat ja VowCH lea soggevokála rievdama báhcahat, mat morfofonologalaš transdusera njuolggadusain leat eaktun (gč. govvosa 4).

```
s s → s          /  _   Vow*  WeG
r j → r j j      /  _   Vow*  WeG
i → á            /  _   VowCh
```

Govus 4. Morfofonologalaš transduser. Konsonántaguovddáža molsašuddan lea «/»-symbola gurutbealde, ja eaktu fas olgešbealde. * mearkkaša «nolla dahje eanet».

Kompilašuvnnain jávkaduvvojit leksikon- ja morfofonologalaš transduseriid gaskadásat ovddasteamit (gč. govvosa 2). Boađus lea ođđa morfologalaš transduser mainna sáhttá sihke analyseret sátnehámiid ja genereret sátnehámiid. Go analysere *gusa* ja *gussii* sátnehámiid, de oažžu leksema ja grammatihkalaš gilkoriid. Jus manná nuppe guvlui, de generere sátnehámiid *gusa* ja *gussii* (gč. govvosa 5).

```
gusa        gussa+N+Sg+Acc
gusa        gussa+N+Sg+Gen
gussii      gussa+N+Sg+Ill
```

Govus 5. Morfologalaš analysáhtor: Olgeš bealde lea atta, gurut bealde fas buvttus.

Jus eai leat dárkilis ráddjejumit, de sáhttá genereret hámiid mat eai gávdno. Dakkár riska ii leat statistihkalaš lahkonanvugiin.

Gillii gullet maiddái goallossánit, ja daid ii leat vejolaš leksikaliseret go gillii jámma ihtet ođđa goallosteamit. Analysáhtoris lea dan dihte dynámalaš substantiiva+substantiiva goallosteapmi, mii maiddái generere goallossániid mat eai gávdno sámegielas, sihke semantihkalačča t ja čállinhámi ektui. Manjit čuovvumuša sáhttá eastadit njuolggadusaiguin ja ráddjejumiiguin (Moshagen – Omma – Pieski 2009).

Goallossáni mearusoassi sáhttá leat ol.nom, ol.gen dahje ml.gen. Vuogádaga ovdaválljen lea ol.nom. Leksikonii lea merkejuvvon dalle go leat eará vejolašvuođat, gč. (1). Goallossáni mearusoasi kasusa sáhttá stivret maiddái vuođđooasi bokte (2):

(1) loddi ; +SgNomCmp SgGenCmp (omd. loddebivdu, lottečivga)
(2) lávlun ; +SgNomLeft +SgGenLeft (omd. sálbmalávlun, sálmmalávlun)

Goallossáni mearusoassi galgá muhtun goallossániin leat allegrohámis go lea ovttaidlogus, eará sániin fas dat ii galgga leat, dahje sáhttá leat. Dát čovdojuvvo leksikonas iešguđetlágan joatkkaleksikonaid (bálgáid) bokte. Goallossániid mat eai heive dán vuogádahkii, sáhttá leksikaliseret.

Giellaoahpalaččat sojahusa ja suorggideami erohus lea ahte sojahusa boađus lea sátnehápmi, ja suorggideami boađus lea ođđa leksema. Juohke substantiivamáddagis manná bálggis kásusmorfologiija čađa, ja máddagii lasihuvvojit kásusgehčosat. Dasa lassin juohke substantiivva ii-nominatiivamátta joatká omd. diminutiiva-vuolleleksikonii. Diminutiiva-leksikon lasiha máddagii gehčosa (-*š*- dahje -*ž*-) ja diminutiivagilkora (+Der/Dimin). Suorggideami boađus sáhttá maiddái leat ođđa sátneluohkká. Ovdamearkka dihte vearbamáddagat sáhttet joatkit Der/eapmi-leksikonii, nu ahte vearbbas

boahtá substantiiva, nugo *vuoruhit + eapmi => vuoruheapmi*. Teknihkalačča suorggidahttin ja sojahus gieđahallojuvvojit seammá vuogi mielde.

Go galgá ráhkadit morfologalaš transdusera, de dárbbaša:

- Sátnelisttu digitálalaš hámis, mas juohke sátnái leat lasihuvvon dieđut mátta- ja sojahanluohká birra.
- Morfofonologalaš, dahje rievtti mielde morfografemihkalaš njuolggadusaid, vai sáhttá generaliseret jeavddalaš geažusmolsašumiid.
- Sorterejuvvon listtu mas leat rájálaš luohkáid sátnehámit ja spiehkastagat rabas sátneluohkáin.
- Buori referánsagrammatihka, gos lea vejolaš gávdnat jeavddalaš ja jeavddahis sojahanminstariid.

Giellatekno/Divvuma vásáhusaid vuođul árvvoštalle ahte morfologalaš transdusera ádjána 0,5–1,5 bargojagi ráhkadit, dan mielde man máŋgabealálaš morfologiija gielas lea. Go vuos lea ráhkadan dakkár analysáhtora, de sáhttá álkit kompileret iešguđetlágan variánttaid, omd. deskriptiiva vs. normatiiva analysáhtora.

2.4. Čoahkkáigeassu

Dán kapihttalis leat čilgejuvvon golbma lahkonanvuogi sátneanalyseremii. Gielaide main lea unnán morfologiija sáhttá geavahit máddagasti diehtoohcamii, ja statistihkalaš lahkonanvugiin ráhkadit sátnedárkkistanprográmma. Sámegielaide ii goabbáge vuohki heive. Dan sajis leat válljen oahpahit grammatihka dihtorii nu ahte sáhttá analyseret ja genereret juohke teorehtalačča vejolaš sojahanhámi. Lea stuora bargu dan dahkat, muhto Kap. 4 čájeha mo grammatihkalaš analysáhtoriin sáhttá ráhkadit máŋga ávkkálaš geavaheaddjiprográmma.

3. Cealkagiid analyseren

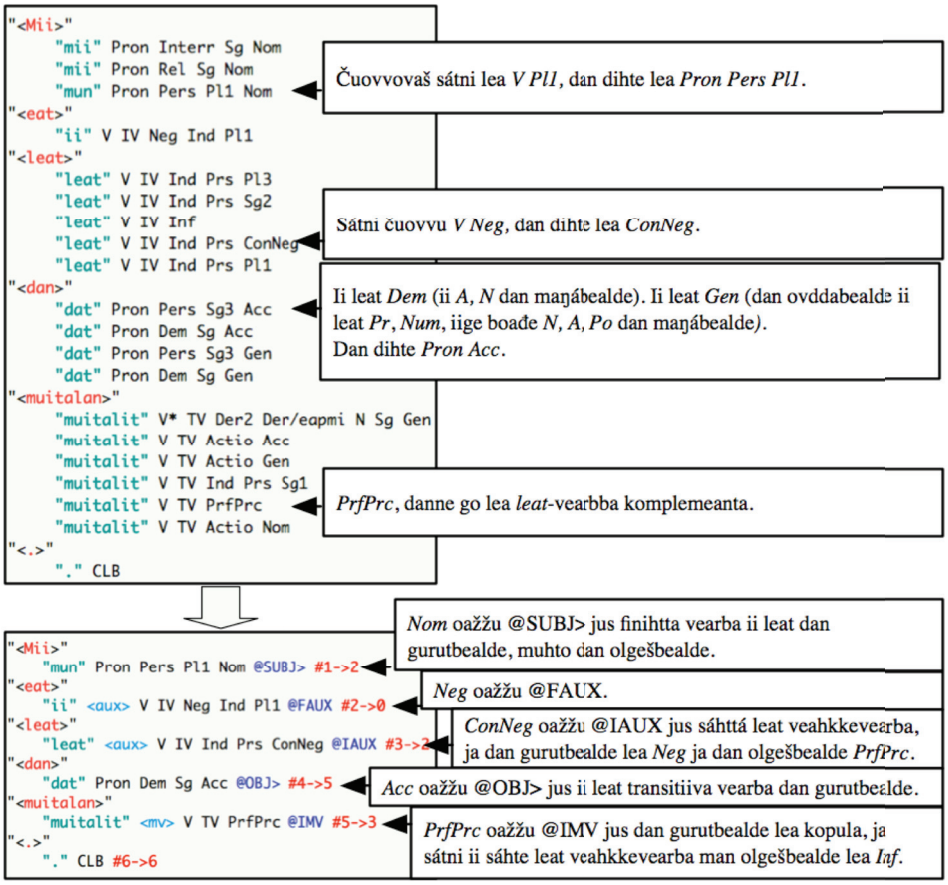
3.1. Homonymiijaid disambigueren

Oktilis teavsttas lea ollu homonymiija. Davvisámegiel sátnehámiin leat gaskamearálaččat 2,7 vejolaš morfologalaš analysa, julevsámegiel sátnehámiin fas 2,2 vejolaš analysa. Davvisámegielas leat eanet homonymat earenoamážit danne go julevsámegiella lea konservatiivvalaččat go davvisámegiella. (Trosterud – Wiechetek 2007.)

Go olmmoš gullá sámegiela, de hárve lea eahpečielggas maid hubmi oaivvilda, danne go sánit gullet man nu kontekstii, ja olbmos leat sihke syntávssalaš ja semánttalaš dieđut. Maiddái automáhtalaš analysa galggašii leat guovttečilggolašvuođaid haga.

Homonymiija sáhttá disambigueret sihke statistihkalaččat ja grammatihkalaččat. Kapihttalis 2 letne ákkastallan manne leat válljen bargat grammatihkalaččat eatge statistihkalaččat. Grammatihkalaččat sáhttá bargat guovtti ládje, badjin–vulos, dahje vuollin–bajás. Badjin–vulos-analysáhtor geahččaladdá hypotesaid das makkár syntávssalaš struktuvra cealkagis lea, ja disambiguere dan mielde. Muhto lunddolaš giella ii álo čuovo syntávssalaš njuolggadusaid; hubmi sáhttá rievdadit oaivila gaskan cealkaga, son sáhttá lasihit guhkes čilgehusaid, dahje cealkka šaddá njulgestaga ilá kompleaksa. Dalle badjin–vulos-analysáhtor ii gávnna makkárge analysa. Maiddái buorit badjin–vulos-analysáhtorat hárve máhttet analyseret eanet go 60 % cealkagiin oktilis teavsttas.

Giellatekno analysáhtor geavaha grammatihkalaš vuollin–bajás-analysáhtora mii nagoda addit analysa maiddái cealkkafragmeanttaide ja kompleaksa cealkagiidda. Gávdnojit máŋgga lágan analysáhtorat, ja Giellatekno analysáhtor vuođđuduvvá ráddjehusgrammatihkkii (*constraint grammar*) (Karlsson – Voutilainen – Anttila 1995; Tapanainen 1996). Grammatihkas leat ráddjehusat (*constraints*) das makkár konteavsttain iešguhtet morfologalaš analysa sáhttá leat. Ráddjehusgrammatihkkii gullet njuolggadusat, ja njuolggadusat kompilerejuvvojit vislcg3-prográmmain (Visl-group 2008). Govus 6 čájeha mo njuolggadusat doibmet, omd. go *mii*-sátnehápmi lea cealkaga álggus ja *eat*-vearbba ovddabealde, de lea rivttes analysa



Govus 6. Ráddjehusnjuolggadusaiguin morfologalaš analysa disambiguerejuvvo ja dasa lassin lasihuvvojit syntáksagilkorat, main lea @, ja dependeansagilkorat: #iežas sajádat → oaivvi sajádat.

Muhtun dáhpáhusain lea dárbbašlaš maiddái čujuhit sániid semánttalaš sisdollui:

(3) *Mun boran eatni gievkkanis*.
(4) *Mun boran láibbi gievkkanis*.

Cealkagis (3) lea *eatni* genetiiva danne go čujuha olbmui, ja cealkagis (4) *láibbi* lea akkusatiiva seammá syntávssalaš konteavsttas. Vai analysáhtor sáhttá válljet rivttes kásusa, de sáhttit lasihit gilkora (<hum>) buot leksemaide mat čujuhit olbmuide, dahje čohkket diekkár leksemaid seahtaide. Ráddjehusgrammatihka njuolggadus čujuha juogo gilkorii dahje sehttii, ja sáhttá dainna lágiin váldit vuhtii sáni semánttalaš sisdoalu.

Davvisámegielas lea akkusatiivvas ja genetiivvas homonymiija olles substantiivaparadigmas, ja opposišuvdna lea dušše lohkosániin. Morfologalaš analysáhtor addá dattetge sihke akkusatiiva- ja genetiivagilkoriid maiddái substantiivvaide. Dán homonymiija lea váttis disambigueret danne go akkusatiiva ja genetiiva sáhttet leat seammá syntávssalaš konteavsttas, nugo bajábeal ovdamearkkas. Máŋgga oktavuođas lea dattetge ávkkálaš earuhit akkusatiivva ja genetiivva, omd. jus dihtor galgá máhttit jorgalit gielas nubbái. Maiddái dalle go automáhtalaččat lasiha syntávssalaš funkšuvdnagilkoriid sániide, lea ávkkálaš earuhit akkusatiivva ja genetiivva.

Davvisámegiela morfologalaš analysáhtor gárvánii vuosttažin, ja dan dihte leatge bargan eanemusat davvisámi disambiguáhtoriin. Dan F-score[5] lea 0,99 sátneluohkkádisambigueremis ja 0,94 olles morfologalaš disambigueremis. Julevsámegiela F-score leat fas 0,95 ja 0,88 (skábmamánu 2009). Eará sámegielaide eat leat vuos ráhkadan disambiguáhtora.

3.2. Syntávssalaš gilkorat

Mii geavahit maiddái ráddjehusgrammatihka merket syntávssa ja dependeanssa gilkoriid sániide. Sámegielas lea viehka friddja sátneortnet, ja mii leat válljen addit olles 50 funkšuvdnagilkora. Omd. lasihuvvojit ieš-guđetlágan subjeaktagilkorat dan mielde leago finihtta vearba olgeš vai gurut bealde, ja leago infinihtta vearbba subjeaktan. Disambiguáhtor ja syntávssalaš analysáhtor leat muhtun muddui ovttastahttojuvvon danne go muhtun syntávssalaš gilkorat leat veahkkin disambigueremiin. Davvisámegiela syntávssalaš analysáhtoriin oažžu F-score 0,93, julevsámegiela analysáhtoriin fas 0,86 (skábmamánu 2009).

Álggus mii ráhkadeimmet sierra syntávssalaš analysáhtora juohke sámegillii, muhto leat ovttastahttigoahtán buot sámegielaid ovtta analysáhtorii. Syntávssalaš erohusat gielaid gaskkas váldojuvvojit vuhtii go lea vejolaš čujuhit morfologalaš analysáhtora buktosa giellagilkorii (<sme> <smj> <sma>). Omd. lullisámegielas kopula ii dárbbaš leat mielde eksistentiála iige habitiiva cealkagis, ja dalle cealkagis ii leat ollenge finihtta vearba. Davvisámegielas

[5] F-score lea vuohki mihtidit dárkilvuođa, ja presišuvnna ja deaivanmeari gaskamearálaš logu. Presišuvnna oažžu go juohká «rivttes analysa» -logu buot bohtosiiguin, ja deaivanmeari go juohká «rivttes analysa» -logu buot analysaiguin maid dihtor livčči galgan gávdnat. Buoremus vejolaš F-score lea 1, heajumus lea 0.

cealkka finihtta vearbba haga analyserejuvvošii fragmeantan dahje elliptalaš cealkkan. Nubbi ovdamearka lea habitiiva kásus mii lea iešguđetlágan gielas gillii – lokatiiva (davvisámegielas), inessiiva (julevsámegielas) ja genetiiva (lullisámegielas). (Antonsen – Trosterud – Wiechetek 2010.)

Min syntávssalaš gilkorat vuođđuduvvojit kompromissii sámi grammatihkalaš árbevieru ja ráddjehusgrammatihka konvenšuvnnaid gaskkas. Váldoerohus lea ahte ráddjehusgrammatihkka lea lineára vuogádat, ja gilkorat addojuvvojit sátnehámiide, eaige gihpuide:

(5) *Mu.*@>N *čeahci.*@SUBJ> *duddjo.*@+FMAINV *čáppa.*@>N *niibbi.*@<OBJ.

Cealkagis (5) dušše *čeahci* oažžu @SUBJ> gilkora (@-mearka earuha syntávssalaš gilkora morfologalaš gilkoriin). *Mu* oažžu @>N, mii muitala ahte sátnehápmi modifisere substantiivva mii lea olgeš bealde. Ja de ferte lohkki dulkot @>N ja @SUBJ> kombinašuvnna gihppun.

Syntávssalaš analysáhtora buktosis sáhttá ráddjehusgrammatihkain dahkat dependeansamuora dependeansagilkoriiguin, nugo govvosis 6. Ruohtas lea merkejuvvon 0. Cealkaga sánit nummarastojuvvojit dađistaga ja njuolla čujuha sáni oaivái, omd. 2→1, mii máksá ahte cealkaga nubbi sátni lea cealkaga vuosttaš sáni dependeanta.

Giellatekno/Divvun lea ráhkadeamen analysáhtoriid sámegielaide. Tabealla 8 čájeha man muttus bargu lei skábmamánus 2009.

Tabealla 8. Sámegielaid analysáhtorat. Dan sajis go ráhkadit sierra syntávssalaš analysáhtora lullisámegiela várás, de leat ovttastahttigoahtán buot sámegielaid ovtta analysáhtorii, mii ii leat vuos gárvvis.

Giella	**Morfologalaš analysáhtor**	**Syntávssalaš analysáhtor**	**Dependeansa-analysáhtor**
Davvisámegiella	101.000 leksema (57.000 leksema)	3450 njuolggadusa – 50 syntávssalaš gilkora	270 njuolggadusa
Julevsámegiella	47.000 leksema	874 njuolggadusa	
Lullisámegiella	30.000 leksema – válmmaš 2010	–	

4. Geavaheaddjiprográmmat

Grammatihkalaš analysáhtor geavahuvvo juo geavaheddjiid prográmmain.

Čállindárkkistanprográmma Divvun (Gaup – Moshagen – Omma – Palismaa – Pieski – Trosterud 2006) vuođđuduvvá morfologalaš automáhtaide mat generererejit buot normatiiva sátnehámiid nugo čilgejuvvo kapihttalis 2.3. Prográmma válbmanii davvisámegillii ja julevsámegillii jagi 2007 loahpageahčen, ja jagi 2010 gárvána maiddái lullisámegiel veršuvdna. Divvun-prográmmas lea funkšuvdna mii geavaheaddjái evttoha riekta čállon sátnehámiid dihto algoritmma mielde.

Sámegielaid siskkáldas morfofonologalaš proseassat dahket oahpahallái váttisin diehtit guđe leksemii sátnehápmi gullá. Dušše 7,9 % sániin davvisámegiel oktilis teavsttas leat seammaláganat go lemmahápmi (Antonsen – Gerstenberger – Moshagen – Trosterud 2009). Dan dihte Giellatekno/Divvun lea morfologalaš generáhtoriin lasihan sihke leksemaid sojahanhámiid ja grammatihkalaš sániid *Vuosttaš Digisánit-* ja *Åarjelsaemien Digibaakoeh*-nammasaš digitálalaš sátnegirjjiide. Sátnegirji sihke muitala guđe grammatihkalaš sánis lea sáhka (gč. govvosa 7), ja addá dieđuid sojahanparadigma birra, nu gohčoduvvon čoavddahámiid, vai geavaheaddji oaidná morfofonologalaš proseassaid mat gullet leksemii. Dákkár sátnegirjji lea álki ráhkadit go gielas lea morfologalaš analysáhtor, ja go gávdno guovttegielalaš sátnelistu maid sáhttá geavahit.

Govus 7. Voestes Digibaakoeh-prográmma dovdá sojahuvvon sániid. Prográmma sáhttá viežžat http://giellatekno.uit.no/words/dicts/index-sma.sme.html

Giellatekno lea ráhkadan giellaoahppanprográmmaid, OAHPA ja VISL (Antonsen – Baal – Trosterud 2009, Antonsen – Huhmarniemi – Trosterud 2009). Morfologalaš ja syntávssalaš analysa dahká vejolažžan ii dušše muitalit leago oahppi vástádus riekta vai ii, muhto maiddái analyseret

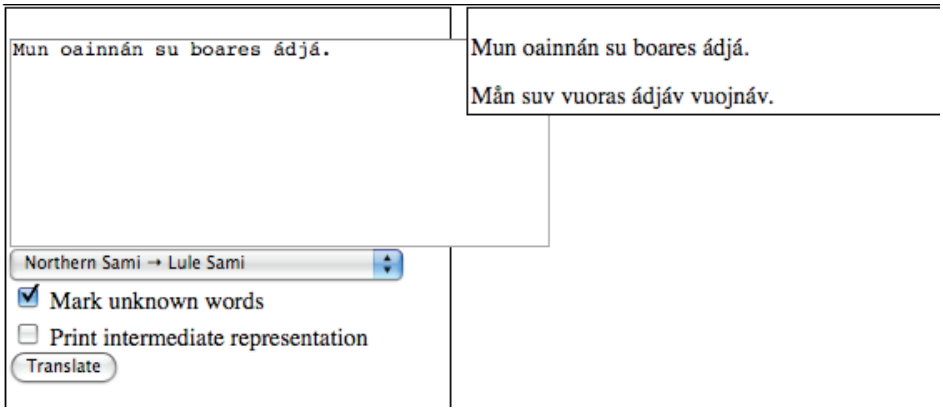
boasttuvuođaid. Guovtti oasseprográmmas (*Vasta* ja *Sahka*) oahppi beassá ieš formuleret vástádusa dihtora jearaldagaide. Dihtora jearaldat ja oahpahalli vástádus addojuvvojit oktan analysáhtorii, mii analysere daid morfologalaččat ja syntávssalaččat. Analysáhtor lasiha guovttelágan gilkoriid, nugo govvosis 8:

- Semánttalaš gilkora mii muitala maid oahppi lea vástidan vai dihtora čuovvovaš jearaldat dahje kommentára heive oahppi vástádussii (&dia-hivsset danne go oahppi lea vástidan ahte TV galgá hivssegii).
- Feailagilkora mii muitala ahte oahppi lea geavahan boasttu sojaheami. Dáinna son oažžu bagadeaddji máhcahusa sierra lásas (&grm-missing-Ill danne go oahppi vástádusas ii leat illatiiva).

```
"<Guđe>"
   "guhte" Pron Interr Sg Gen &grm-missing-Ill
"<latnjii>"
   "latnja" N Sg Ill
"<moai>"
   "mun" Pron Pers Du1 Nom
"<bidje>"
   "bidjat" V TV Ind Prs Du1
"<mu>"
   "mun" Pron Pers Sg1 Gen
"<TV>"
   "TV" N ACR Sg Acc
"<^qst>"
   "^sahka" QDL gosa_bidjat_TV &dia-hivsset
"<Moai>"
   "mun" Pron Pers Du1 Nom
"<bidje>"
   "bidjat" V TV Ind Prs Du1
"<TV>"
   "TV" N ACR Sg Gen
"<hivssegis>"
   "hivsset" N Sg Loc
"<.>"
   "." CLB
```

Govus 8. Sahka-prográmma lasiha jearaldat/vástádus-bárrii sihke navigeren- ja grammatihkkagilkora. Gilkoriid vuođul vuogádat addá máhcahusa oahppái. Prográmma lea interneahtas: http://oahpa.uit.no

Mášenjorgaleamis geavahuvvo eanaš statistihkalaš lahkonanvuohki, dego Google Translate-prográmmas. Dakkár teknologiija ii heive sámegillii, go dárbbaša stuorát teakstačoakkáldagaid go mat davvisámegillii gávdnojit. Sámegillii lea baicca lunddolaš geavahit grammatihkalaš analysáhtoriid. Mášenjorgalanprográmma Apertium lea ráhkaduvvon jorgalit fuolkegielaid gaskkas (nugo katalána- ja spánskkagiela gaskkas). Apertium-prográmma doaibmá rabas gáldokodaprinsihpa mielde (http://wiki.apertium.org), ja dihtorlingvisttat olles máilmmis oassálastet bargui lasihit ođđa giellapáraid dasa. Dál leat 35 doaibmi giellabára, ja 27 jorgalanbára leat álggahuvvon. Daidda gullá prošeakta jorgalit davvisámegielas julevsámegillii (Tyers – Wiechetek – Trosterud 2009) (gč. govvosa 9).



Govus 9. Apertium-prográmmain lea vejolaš jorgalit davvisámegielas julevsámegillii. Prográmma lea interneahtas: http://xixona.dlsi.ua.es/testing/

5. Servodatlaš váikkuhusat

5.1. Giellateknologiija servodagas

Dábáleamos giellateknologalaš heiveheapmi lea dokumeantaohcan. Dokumeanttaid siskkáldas vuoruheapmi lea mearriduvvon liŋkastruktuvrra bokte: dat dokumeanta vuoitá masa čujuhit máŋga liŋkka. Ovdalgo vuoruha, de ferte gávdnat áššáiguoskevaš dokumeanttaid. Sámegiela oktavuođas morfologalaš variašuvdna lea dehálaš faktor: jos ohcá sáni *giella*, de giellateknologiija haga ii gávnna sániid *gielas, gillii, giellan*. Máddagastima bokte maŋimus sátnehámi goitge gávdná (*-n* váldojuvvo

eret). Davvisámegiela máddagastima bokte substantiivvain leat gaskamearálaččat 3,1 máddaga (gč. kap. 2.1), nu ahte 2/3 vejolaš relevánta sátnehámiin leat oaidnemeahttumat. Morfologalaš analysa livčči dan dihte dehálaš dokumeantaohcamii.

Buriid giellateknologalaš resurssaid haga ii oktage giella sáhte boahtteáiggis:

- doaibmat hálddahuslaš giellan. (Teaksta ferte dárkkistuvvot, máŋggalágan skovit ráhkaduvvot, olbmot dárbbašit čoahkkáigeasuid. Buot dát galgá dahkkot automáhtalaččat, giellateknologalaš reaidduiguin.)
- doaibmat guovttegielalaš hálddahusas. (Máŋggagielalaš veršuvnnat seamma teavsttas ráhkaduvvojit automáhtalaččat mášenjorgaleami bokte. Teavsttain dárkkistuvvo automáhtalaččat ahte tearbmageavaheapmi lea konsekveanta.)
- vurket dokumeanttaid digitála arkiivvaide. (Gáldomášiinnat geavahit giellateknologiija sihke dokumeanttaid ja dieđuid klassifiseremii ja maiddái vurkejuvvon teavsttaid gávdnamii. Livčči maiddái vuogas jus sáhtášii ohcat dihto sisdoalu beroškeahttá das guđe gillii teaksta lea čállojuvvon.)

Go oaidná vejolašvuođaid maid dihtor attášii, de olmmoš ákkastalašii geavahit dárogiela ja eará eanetlogu giela dakkár sajis masa ii gávdno giellateknologiija. Manne galgá vurket dokumeantta dakkár čállojuvvon gillii mii dahká dan veadjemeahttumin fas gávdnat?

Maŋimuš logi jagi leat dáhpáhuvvan stuora ovdáneamit suorggis mas leat omd. semánttalaš fierpmádagat ja sániid mearkkašumiid (*word-sense*) disambigueren[6]. Dát bargu ii leat vuos ihtán dábálaš geavaheaddji dihtorii, muhto dat boahtá, ja jus mii eat daga maidege, de dat boahtá dušše stuora gielaide. Dađistaga go gávppálaš giellateknologalaš prográmmat buorránit, ja go almmolaš hálddahus vuođđuduvvá giellateknologiijai, de stuorru erohus gielaid gaskkas main leat ja main eai leat dakkár resurssat.

Dáidda geavatlaš čuovvumušaide lassin, de mii lingvisttat maiddái sáhttit

---

[6] Omd. sánis fierbmi leat máŋga mearkkašumi, ja go disambiguere, de gávdná guđe mearkkašumis lea sáhka dihto konteavsttas.

ávkkástallat resurssaiguin: Bures ovdánan giellateknologiija dagašii vejolažžan guorahallat giela ollu beaktileappot go giellateknologiija haga. Ja de lea vel duođaid dehálaš sivva: Eanaš gielaide máilmmis ii leat gávppálaččat gánnáhahtti dahkat vuođđobarggu mii dárbbašuvvo giellateknologalaš geavaheaddjiprográmmaide. Bargu báhcá akademalaš dutkanásahusaid vuoruheami duohkai. Muhto lingvistii leage mávssolaš bargu čállit ollislaš referánsagrammatihka nu ollu gielaide go vejolaš.

5.2. Olaheamit ja ođđa hástalusat

Sámegielaide stuorimus servodatlaš váikkuhus dán rádjái lea leamaš dihtora vuogádaga ja boallobeavddi lokaliseren. Unicode-standárdda, Sámi dihtorlávdegotti ja Ruoŧa standardiserenlávdegotti ánsun juohke dihtoris leat dál sámegiel boallobeavdi ja bustávat.

Čuovvovaš dehálaš boađus lei Divvun-prográmma, mii ee. lea dahkan vejolažžan buvttadit sámegiel aviissaid jođáneappot (Guhkes lávki ovddasguvlui 2007), ja maid Ávvira váldodoaimmaheaddji atná buoremus veahkkeneavvun maid sámegielat aviisa lea ožžon ođđa áiggis (Mii dárbbašit «čuorbbi» 2008).

Boahtteáiggis máŋggagielalašvuohta šaddá dehálaččat aht' dehálaččat. Norgga stáhtahálddahusa ođđa giellapolitihka mielde Norgga stáhta lea interneahtas njealjegielat (guokte dárogiela, eŋgelasgiella, davvisámegiella)[7]. Dakkár politihka lea guhkit áigeperspektiivvas vejolaš čađahit dušše mášenjorgalemiin.

Sámi ásahusain ja servviin lea muhtumin ággan čállit dárogillii dat ahte eai buot lohkkit hálddaš sámegiela. Ja nu sámegiella dávjá báhcá jorgalusgiellan, dan muddui go lea áigi ja resurssat dasa. Mášenjorgalanprográmma sámegielas dárogillii dagašii vejolažžan čállit njuolgga sámegillii, go prográmma veahkehivčči olbmuid geat eai ipmir sámegiela.

Stuora hástalus lea jorgalit doarvái teavsttaid julevsámegillii ja lullisámegillii. Sámelága giellanjuolggadusaid ollašuhttimii eai gávdno doarvái čállit

[7] Geahča omd. http://ráđđehus.no/. Neahttasiiddus leat njeallje giellamolssaeavttu: Bokmål, Nynorsk, Sámegiella, English.

eaige jorgaleaddjit julev- ja lullisámegillii. Sáhttá leat váttis geargat buot skuvlagirjjiid jorgalahttit ovdalgo čuovvovaš oahppoplánaođastus boahtá. Buorre veahkkin dan bargui livčči mášenjorgaleapmi davvisámegiela bokte.

6. Konklušuvdna

Buot digitálalaš sámegiela teakstagieđahallan vuođđuduvvá dasa ahte lea vejolaš čállit ja lohkat sámegiela digitálalaš hámis. Maŋimuš logi jagi lea dát leamaš vejolaš buot sámegielaide, muhto eai buot almmolaš registarat leat atnigoahtán ođđa teknologiija, nu ahte sámegielat dieđuid ii velge leat vejolaš registreret omd. Brønnøysundregistariidda iige Álbmotregistarii.

Maŋimuš jagiid leat boahtán syntávssalaš ja morfologalaš analysaprográmmat davvi-, julev- ja lullisámegillii. Nugo čájehuvvon 2. kapihttalis, de sámegielaid ii sáhte analyseret seammá lahkonanvugiiguin go gielaid main ii leat nu rikkis morfologiija. Sámegielaide grammatihkalaš lahkonanvuohki lea dárbbašlaš. Analysaprográmmat barget guovtti proseassas, nubbi lasiha gehčosiid sáni máddagii ja nubbi dahká morfofonologalaš proseassaid. Boađus lea prográmma mii sáhttá analyseret ja genereret juohke sámegiel sátnehámi. Syntávssalaš analyserenprográmma vállje rivttes morfologalaš analysa, addá syntávssalaš funkšuvnna ja dependeansaanalysa dan vuođul makkár relatiiva sajádat sátnehámis lea cealkagis.

Analysáhtorat veahkehit min ipmirdit sámi grammatihka, muhto dain leat maiddái dát dehálaš doaimmat: Daiguin sáhttá analyseret stuora teakstačoakkáldagaid vai gielladutkit sáhttet generaliseret. Ja daid vuođul lea vejolaš ráhkadit prográmmaid mat leat dehálaččat buot sámegielagiidda: Divvunprográmmaid, interaktiivalaš sátnegirjjiid, giellaoahppanprográmmaid ja dieđuidgieđahallanprográmmaid.

Dáin bohtosiin leat servodatlaš čuovvumušat. Jus sámegiella galgá leat mielde ođđa máŋggagielalaš servodagas, ja jus davvisámegiella ja eará sámegielat galget sáhttit doaibmat hálddahusgiellan ja oahpahusgiellan, de sámegielain fertejit leat seammá buorit veahkkeneavvut go eará gielain. Lea ain ollu bargu dan muddui ollet, muhto vuođđobargu lea juo dahkkon – dihtor máhttá sámegiela grammatihka.

Materiála

DNT = (2005) Det nye testamentet. – http://www.bibelen.no

NT = (1611) New Testament, King James Bible. – http://www.kingjamesbibleonline.org

NAČ = (1994) Finnmárkku eatnamiid ja čázádagaid geavaheapmi historjjálaš geahččamiin. Norgga almmolaš čielggadeamit 1994:21 S. – http://www.regjeringen.no/se/dep/krd/Dokumeanttat/NA-at/1993/nac_199421.html?id=139744

OT = (1998) Ođđa testameanta. – http://www.bibelen.no

Girjjálašvuohta

Antonsen, Lene – Baal, Berit Ánne Bals – Huhmarniemi, Saara – Trosterud, Trond 2009: Dihtor ja giela válljenvejolašvuođat – gielalaš ja pedagogalaš čuolmmat. – Johanna Ijäs – Nils Øivind Helander (doaim.), *Sáhkavuoruin sáhkan. Sámegiela ja sámi girjjálašvuođa muhtin áigeguovdilis dutkanfáttát* s. 87–102. Dieđut 1/2009. Guovdageaidnu: Sámi allaskuvla.

Antonsen, Lene – Gerstenberger, Ciprian-Virgil – Moshagen, Sjur Nørstebø – Trosterud, Trond 2009: Ei intelligent elektronisk ordbok for samisk. – *LexicoNordica* Volum 16 s. 271–283. Oslo: Nordisk forening for leksikografi.

Antonsen, Lene – Huhmarniemi, Saara – Trosterud, Trond 2009: Interactive pedagogical programs based on constraint grammar. *Proceedings of the 17th Nordic Conference of Computational Linguistics.* Nealt Proceedings Series 4. Tartu: Tartu University Library. – http://hdl.handle.net/10062/9546 (16.04.10).

Antonsen, Lene – Trosterud, Trond – Wiechetek, Linda (2010): Reusing Grammatical Resources for New Languages. *Proceedings of the International conference on Language Resources and Evaluation LREC* 2010. Stroudsburg: The Association for Computational Linguistics. – http://www.lrec-conf.org/proceedings/lrec2010/pdf/254_Paper.pdf (10.06.10).

Beesley, Kenneth R. – Karttunen, Lauri 2003: *Finite State Morphology*. Stanford, California: CSLI publications in Computational Linguistics.

Gaup, Børre – Moshagen, Sjur N. – Omma, Thomas – Palismaa, Maren –

Pieski, Tomi – Trosterud, Trond 2006: From Xerox to Aspell: A First Prototype of a North Sámi Speller Based on TWOL Technology. – Anssi Yli-Jyrä – Lauri Karttunen – J. Karhumäki (doaim.), *Finite-State Methods and Natural Language Processing.* Lecture Notes in Computer Science 4002, s. 306–307. Berlin – Heidelberg: Springer-Verlag. – http://www.springerlink.com/content/an651qt0g45k55u1/ (16.04.10).

Guhkes lávki ovddasguvlui [váldočála]. 2007. – *Min Áigi,* 2.6.2007, Nr 42, s. 2.

Karlsson, Fred – Voutilainen, Arto – Heikkilä, Juha – Anttila, Arto 1995: *Constraint grammar. A language-independent system for parsing unrestricted text.* Berlin – New York: Mouton de Gruyter.

Mii dárbbašit «čuorbbi» [váldočála]. 2008. – *Ávvir,* 15.11.2008, Nr 144, s. 2.

Moshagen, Sjur – Omma, Thomas – Pieski, Tomi 2008: *Goallosteapmi Divvun-reaidduin.* Tromsø: Universitetet i Tromsø. – http://giellatekno.uit.no/background/Goallosteapmi_Divvun.pdf (30.11.2009).

Moshagen, Sjur – Sammallahti, Pekka – Trosterud, Trond 2004: Twol at work. – Antti Arppe – Lauri Carlson – Krister Lindén – Jussi Piitulainen – Mickael Suominen – Martti Vainio – Hanna Westerlund – Anssi Yli-Jyrä (doaim.), *Inquiries into Words, Constraints and Contexts* s. 94–105. Stanford, California: CSLI.

Tapanainen, Pasi 1996: The Constraint Grammar Parser CG-2. *Publications of the Department of General Linguistics*, 27. Helsinki: University of Helsinki.

Trosterud, Trond – Wiechetek, Linda 2007: Disambiguering av homonymi i nord- og lulesamisk. – Jussi Ylikoski – Ante Aikio (doaim.), *Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007* s. 401–421. Suomalais-Ugrilaisen Seuran Toimituksia 253. Helsinki: Suomalais-Ugrilainen Seura. – http://www.sgr.fi/sust/sust253/sust253_trosterudjawiechetek.pdf (16.04.10).

Tyers, Francis M. – Wiechetek, Linda – Trosterud, Trond 2009: Developing Prototypes for Machine Translation between Two Sámi Languages. *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09.* Allschwil: European Association for Machine Translation. – http://www.mt-archive.info/EAMT-2009-Tyers-1.pdf (16.04.10).

Visl-Group 2008: *Constraint Grammar*. Odense: University of Southern Denmark. – http://beta.visl.sdu.dk/constraint grammar.html (30.11.09).

Why the computer should know its Sami grammar

Language technology constitutes the foundation for the necessary infrastructure needed for any language to function in a modern literary society. The Sami languages differ from the languages for which most such technology is developed in two important ways: The body of text available (either Sami or bilingual Sami – majority language) constitutes but a fraction of what is available for Western European state languages, and the Sami languages have morphological structures far more complex than the ones for most of the Western European state languages.

The article argues that the answer to this challenge is to build a grammar-based language technology for the Sami languages, and presents ongoing work fulfilling this goal. It is shown how morphophonological processes and inflectional and derivational morphology may be modelled as finite-state transducers, and combined with a syntactic component consisting of context-sensitive constraint grammar rules, to constitute a robust grammatical analyser capable of both analysing running text, and generating any word form. The speech communities of the Sami languages are not large enough to uphold a language technology industry, but the grammar-based language model is interesting for theoretical linguists as well.

Practical applications derived from the basic grammatical analysers include spell-checkers, interactive computer-assisted language learning programs, and machine translation.

Lene Antonsen – Trond Trosterud
Romssa universitehta / University of Tromsø
lene.antonsen@uit.no
trond.trosterud@uit.no