Sigrún Helgadóttir

Markari fyrir íslenskan texta

Sigrún Helgadóttir

Markari fyrir íslenskan texta

Inngangur

Starfshópur sem samdi skýrslu um tungutækni á vegum menntamálaráðuneytisins veturinn 1998-1999 lagði m.a. til að „unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði".

Það að greina orð eftir orðflokkum og beygingu er kallað *tagging* í ensku og greiningarstrengurinn kallast *tag*. Forrit eða kerfi sem framkvæmir þetta verk kallast á ensku *tagger*. Lagt er til að greiningarstrengurinn kallist **mark** á íslensku, aðgerðin verði köllluð **mörkun** og forritið eða kerfið kallist **markari**.

Í anda tillögunnar var gerð málfræðilegs markara fyrir íslensku eitt af þeim verkefnum sem voru styrkt af tungutækniverkefni menntamálaráðuneytisins í apríl 2002. Markmið verkefnisins var að finna aðferðir til þess að greina íslenskan texta vélrænt í orðflokka og eftir beygingu. Afurð verkefnisins átti að vera annaðhvort reglusafn til að nota með svokölluðum Brill-markara eða sérstakt forrit. Verkefnið þróaðist þó á þann veg að prófaðar voru fjórar aðferðir við mörkun íslensks texta. Einnig var prófað að setja þrjár af þessum aðferðum saman eftir tilteknum reglum til þess að ná sem bestum árangri við mörkun.

Í ýmsum tungutækniverkefnum þar sem unnið er úr texta er ávinningur að því að orð í textanum séu greind í orðflokka og beygingarmyndir. Má þar nefna greiningu texta í setningahluta (e. *partial parsing*), orðtöku úr texta fyrir gerð orðasafns, upplýsingaheimt, talkennsl, talgervingu, vélrænar þýðingar, orðabókargerð, fyrirspurnarkerfi og leiðréttingarforrit. Einnig er nauðsynlegt að orð í texta séu greind eftir orðflokkum og beygingu ef gera á tíðnikönnun á texta eins og þá sem birt er í Íslenskri orðtíðnibók (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991).

Handvirk greining texta eftir orðflokkum og beygingu er mjög tímafrek og heldur leiðinleg iðja. Þess vegna hefur lengi verið fengist við að þróa vélrænar aðferðir við það verkefni. Þetta svið hefur því fengið mikla umfjöllun hjá þeim sem vinna við máltækni.

Vélrænar aðferðir við mörkun eru venjulega flokkaðar í tvo flokka, regluaðferðir og tölfræðilegar aðferðir. Fyrstu vélrænu aðferðirnar sem var beitt voru regluaðferðir. Orðasafn var notað til að merkja sérhvert orð í texta með öllum hugsanlegum greiningarstrengjum. Síðan voru notaðar reglur til þess að skera úr um hvaða greiningarstrengur væri réttur. Þessar reglur voru byggðar á málfræði hvers tungumáls og venjulega samdar af málfræðingum.

Tölfræðilegar aðferðir byggjast allar á því að orðum í textasafni hefur verið úthlutað mörkum og þau leiðrétt handvirkt (*data-driven methods*). Forrit er síðan látið búa til líkan á grundvelli þessara gagna sem þegar hafa verið greind. Aðrar aðferðir sem ekki eru beinlínis flokkaðar sem tölfræðilegar byggjast einnig á sama vinnulagi. Má þar nefna aðferð sem mætti kalla leiðréttingaaðferð og byggist á því að skipta um greiningarstreng þegar ákveðnum skilyrðum í umhverfi orðsins er fullnægt (e. *transformation-based learning*). Forrit eða kerfi sem nota fyrir fram greint textasafn til þess að læra af mætti kalla námfúsa markara.

Markmið verkefnisins var að búa til markara sem gæti markað íslenskan texta með a.m.k. 92% nákvæmni.

Efniviður og aðferðir

Prófaðar voru nokkrar aðferðir við mörkun sem allar eiga það sameiginlegt að reynt er að læra af fyrir fram greindum gögnum. Námfús markari lærir fyrst af textasafni sem hefur verið greint í orðflokka og eftir beygingu. Markarinn nýtir síðan þessa kunnáttu til þess að marka orð í texta sem hann hefur ekki lesið áður. Til þess að ná sem bestum árangri er æskilegt að textinn sem á að marka sé sem líkastur textanum sem markarinn lærði af.

Til þess að prófa námfúsan markara á nýju tungumáli eða nýrri gerð af texta er nauðsynlegt að hafa aðgang að stóru textasafni þar sem hvert orð hefur verið greint eftir orðflokki og beygingu. Textasafninu er venjulega skipt í tvo hluta. Annar hlutinn, sem gæti verið um 90% af safninu, er notaður til þess að þjálfa markarann og kallast þjálfunarsafn. Hinn hlutinn (10% af safninu) er notaður til þess að prófa það sem markarinn hefur lært og kallast prófunarsafn. Orðum er úthlutað marki og útkoman síðan borin saman við rétt mark.

Í þeirri vinnu sem hér er greint frá var notað textasafn sem varð til við undirbúning *Íslenskrar orðtíðnibókar*. Í textasafninu eru 590.297 lesmálsorð sem birtast í 59.358 mismunandi orðmyndum að meðtöldum greinarmerkjum. Lesmálsorðunum fylgja 639 mismunandi greiningarstrengir að meðtöldum greinarmerkjum.

Ákveðið var að prófa fjórar aðferðir við mörkun. Tvær þeirra teljast til tölfræðilegra aðferða, eina mætti kalla leiðréttingaaðferð (e. *error-driven transformation-based learning*) og ein byggist á minnistækni (e. *memory-based technique*). Alls voru próf-

aðir fimm markarar sem unnt er að þjálfa á íslenskum texta og eru fáanlegir endurgjaldslaust. Tölfræðimarkararnir sem voru prófaðir voru **TnT**, sem byggist á Markovslíkani, og **MXPOST** sem byggist á svo kölluðu hámarksóreiðulíkani (e. *Maximum Entropy Model*). Tveir markarar, **µ-TBL** og **fnTBL**, sem byggjast á leiðréttingaaðferðinni voru prófaðir og einn markari, **MBT**, sem byggist á minnistækni.

Prófanir

Tölvuskrár Orðtíðnibókarinnar eru skipulagðar þannig að í hverri skrá er textabútur úr einni heimild. Hverri skrá var skipt í 10 nokkurn veginn jafna búta. Úr þessum 10 bútum voru búin til 10 pör af skrám þannig að skrárnar í hverju pari skarast ekki. Í hverju pari er ein skrá með um 90% af lesmálsorðum úr textasafninu og önnur með um 10% af lesmálsorðum úr textasafninu. Stærri skráin er notuð sem þjálfunarsafn og sú minni sem prófunarsafn. Í hverju pari eru því textar sem eiga að vera dæmigerðir fyrir alla textaflokka í textasafninu. Prófunarsöfnin 10 eru óháð hvert öðru en þjálfunarsöfnin hafa um 80% sameiginlega texta. Allir markarar voru prófaðir á öllum 10 pörum og fundin meðalnákvæmni (*ten-fold cross-validation*).

Tveir markaranna, µ-TBL og MBT, virtust ekki henta fyrir íslenska textasafnið en markararnir TnT, MXPOST og fnTBL voru prófaðir á öllum 10 pörunum. Eins og sést af töflu 1 gaf TnT-markarinn besta niðurstöðu, þá MXPOST-markarinn og sístur var fnTBL-markarinn.

Nákvæmni er sýnd fyrir öll orð, þekkt orð og óþekkt orð. Óþekkt orð eru orð sem eru í prófunarsafni en ekki viðkomandi þjálfunarsafni. Tölur í töflu 1 eru fengnar með því að leggja saman prófunarsöfnin 10 og telja rétt greind orð fyrir hverja mörkunaraðferð. Tölur um nákvæmni gefa því meðalnákvæmni fyrir pörin 10. Meðalhlutfall óþekktra orða í prófunarsöfnunum var 6,84%.

Eins og sést á töflunni eru markararnir þrír misjafnlega duglegir við að greina óþekkt orð, þ.e. orð sem þeir hafa ekki séð áður. Markararnir nota mismunandi aðferðir við greiningu óþekktra orða. TnT-markarinn virðist hafa yfir að ráða betri aðferð en hinir markararnir við að greina óþekkt orð og fær því besta heildarniðurstöðu eða **90,36**%.

Vert er að benda á að mark er talið rangt þó að aðeins eitt af sex atriðum í greiningarstreng sé rangt.

Niðurstöður mörkunar voru skoðaðar nákvæmlega og greindar til þess að finna hvers konar villur markararnir gera og hvernig mætti bæta árangurinn. Algengustu villurnar sem allir markarar gera er að rugla saman fallstjórn forsetninga. Næst kemur ruglingur á milli beygingarmynda nafnorða sem hafa sömu mynd. Má þar nefna þolfall og þágufall kvenkynsorða í eintölu (`þf`. *konu;* `þgf`. *konu*) og nefnifall og þolfall hvorugkynsorða í eintölu (`nf`.

barn; þf. *barn*). Ruglingur á milli fyrstu persónu og þriðju persónu eintölu af sögnum er líka algengur þar sem þessar beygingarmyndir líta eins út (*ég fer; hann fer*). Einnig má nefna nafnhátt og þriðju persónu fleirtölu í nútíð en þessar beygingarmyndir líta eins út (*að fara; þeir fara*).

Markararnir gera að nokkru leyti ólíkar villur og það má nota á ýmsa vegu til þess að bæta árangur mörkunar. Prófað var að kjósa á milli marka sem markarar úthluta. Í íslenska verkefninu voru prófaðir þrír markarar. Sú aðferð við kosningu á milli þeirra sem gaf besta niðurstöðu fólst í því að velja það mark sem tveir eða fleiri voru sammála um. Ef allir þrír eru ósammála var valið mark þess markara sem stóð sig best, í þessu tilviki mark TnT. Í töflu 1 sést að með því að kjósa á milli markara á þennan hátt fæst **91,54**% nákvæmni.

Greining textans í textasafni orðtíðnibókarinnar er mun ítarlegri en tíðkast í tungumálum sem þessi kerfi höfðu verið prófuð á. Skrá yfir alla greiningarstrengi eða mörk sem koma fyrir í tilteknu mörkuðu textasafni er oft kölluð **markaskrá** (e. *tagset*). Markaskrá orðtíðnibókarinnar er mjög stór og ítarleg. Sú greining sem þar er notuð er ekki endilega sú eina rétta og verið getur að sumar tungutæknilausnir geti nýtt sér greiningu sem er ekki jafn ítarleg. Sum tungutækniverkefni gætu þurft mikla nákvæmni í mörkun en ekki mjög ítarlega greiningu. Í viðauka sést hvernig greiningarstrengir Orðtíðnibókarinnar eru settir saman.

Tafla 1. Nákvæmni þriggja markara og nákvæmni sem fæst með því að kjósa á milli niðurstöðu markaranna og nákvæmni þegar greiningarstrengir eru einfaldaðir. Að lokum er beitt reglum.
Einnig niðurstöður miðað við að nota orðasafn.
Einnig niðurstöður miðað við að aðeins orðflokkar séu greindir

Markaskrá	Orða-safn[1]	Aðferð	Óþekkt orð		Þekkt orð		Öll orð	
			Tíðni	%	Tíðni	%	Tíðni	%
Alls			40.392	6,84	549.905	93,16	590.297	100,00
Óbreytt	Nei	MXPOST	25.246	62,50	500.617	91,04	525.863	89,08
Óbreytt	Nei	ifdTBL	21.823	54,03	502.378	91,36	524.201	88,80
Óbreytt	Nei	TnT	28.919	71,60	504.484	91,74	533.403	**90,36**
Óbreytt		Kosið milli m	29.003	71,80	511.348	92,99	540.351	**91,54**
Einfölduð[2]		MXPOST	25.255	62,52	508.753	92,52	534.008	90,46
Einfölduð[2]		ifdTBL	21.831	54,05	509.309	92,62	531.140	89,98
Einfölduð[2]		TnT	28.925	71,61	513.173	93,32	542.098	**91,83**
Einf. f. kosn.		Kosið milli m	29.010	71,82	517.342	94,08	546.352	**92,56**
		Mark MXPOS	29.141	72,15	518.775	94,34	547.916	**92,82**
Óbreytt	Nei	MXPOST	25.252	62,50	500.611	91,04	525.863	89,08
Óbreytt	Já	ifdTBL	28.461	70,44	503.142	91,50	531.603	90,06
Óbreytt	Já	TnT	34.859	86,28	505.511	91,93	540.370	**91,54**
Óbreytt		Kosið milli m	34.331	84,97	512.044	93,12	546.375	92,56
Einfölduð[2]		MXPOST	25.261	62,52	508.747	92,52	534.008	90,46
Einfölduð[2]		ifdTBL	28.467	70,46	509.788	92,71	538.255	91,18
Einfölduð[2]		TnT	34.863	86,29	513.797	93,44	548.660	**92,95**
Einf. f. kosn		Kosið milli m	34.336	84,98	517.773	94,16	552.109	**93,53**
		Mark MXPOS	34.013	84,18	518.818	94,35	552.831	**93,65**
Orðfl greindir		MXPOST	35.621	88,19	538.697	97,96	574.318	97,29
		ifdTBL	33.181	82,15	540.859	98,35	574.040	97,25
		TnT	37.349	92,47	541.946	98,55	579.295	98,14

[1] Orðasafn hefur u.þ.b. helming óþekktra orða
[2] Einföldun felst í að greina ekki atviksorð og ekki heldur samtengingar
Fornöfn eru sett í einn flokk en að öðru leyti er greining þeirra eftir kyni, tölu og falli látin haldast.

Prófað var að einfalda greiningarstrengi á þrennan hátt. Einföldunin felst í því að líta aðeins á fyrsta staf í greiningarstreng fyrir atviksorð og samtengingar, þ.e. greina þessa orðflokka ekki í undirflokka, og slá saman fornafnaflokkum en láta greiningu fornafna halda sér að öðru leyti. Í töflu 1 sést að TnT-markarinn nær 91,83% nákvæmni eftir að markaskrá hefur verið einfölduð.

Eins og þegar er getið skiptir miklu máli að hafa góðar aðferðir til þess að greina óþekkt orð. Markararnir búa til orðasafn úr þjálfunarsafninu og nota það við mörkun texta sem þeir hafa ekki séð. Tveir af mörkurunum, TnT og fnTBL, gefa kost á að nota viðbótarorðasafn við mörkunina. Notað var orðasafn sem hefur um helming þeirra orða sem eru óþekkt í hverju prófunarsafni miðað við samsvarandi þjálfunarsafn og bætti það árangur nokkuð. Í töflu 1 sést að TnT-markarinn nær 91,54% nákvæmni með því orðasafni.

Flestir markararnir eiga í erfiðleikum með að greina á milli orðmynda sem líta eins út. Má þar nefna þolfall og þágufall eintölu kvenkynsorða og nefnifall og þolfall eintölu hvorugkynsorða. MXPOST-markarinn virtist gera færri slíkar villur en hinir markararnir tveir. Samdar voru reglur til þess að velja mark MXPOST-markarans frekar en niðurstöðu kosningar ef tilteknum skilyrðum var fullnægt. Með því móti mátti bæta niðurstöðu nokkuð. Tafla 1 sýnir helstu niðurstöður og að besta niðurstaðan fyrir þann efnivið sem var notaður varð **93,65**% nákvæmni.

Neðst í töflunni sést nákvæmni ef aðeins er litið á greiningu eftir orðflokkum. Þá nær TnT 98,14% nákvæmni, MXPOST 97,27% nákvæmni og fnTBL 97,25% nákvæmni. Í sumum tungutækniverkefnum nægir greining eftir orðflokkum.

Aðferðirnar prófaðar á nýjum textum

Aðferðirnar við mörkun sem hér hefur verið lýst voru prófaðar á textum sem ekki voru hluti af textasafni Orðtíðnibókarinnar. Fjögur aðskilin lítil textasöfn voru notuð. Í fyrsta safninu eru brot úr 13 skáldritum frá 19. öld og fyrri hluta 20. aldar, samtals 6.022 lesmálsorð að meðtöldum greinarmerkjum. Í öðru safninu eru brot úr níu skáldverkum frá því eftir 1980, samtals 3.601 lesmálsorð að meðtöldum greinarmerkjum. Í þriðja safninu eru textar um tölvur og tækni sem eru fengnir úr gagnasafni Morgunblaðsins, úr Fréttabréfi RHÍ og af vefsíðum ýmissa tölvufyrirtækja, samtals 2.926 lesmálorð að meðtöldum greinarmerkjum. Í fjórða safninu eru textar um lögfræði og viðskipti sem eru teknir úr Lagasafni, fréttabréfi fjármálaráðuneytis og Morgunblaðinu (viðskipti), alls 2.776 lesmálsorð að meðtöldum greinarmerkjum. Mörkun var síðan leiðrétt til þess að unnt væri að reikna út nákvæmni mörkunar með hinum ýmsu aðferðum.

Tafla 2. Nákvæmni við mörkun texta sem eru ekki í textasafni Orðtíðnibókar

Gamall bókmenntatexti						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	524	8,70	5.498	91,30	6.022	100,00
MXP	334	63,74	4.935	89,76	5.269	87,50
fnTBL	279	53,24	4.985	90,67	5.264	87,41
TnT	393	75,00	5.209	94,74	5.602	93,03
TnT, einf.	393	75,00	5.218	94,91	5.611	93,18
MXP, orðfl.	458	87,40	5.326	96,87	5.784	96,05
fnTBL, orðfl	409	78,05	5.374	97,74	5.783	96,03
TnT, orðfl.	472	90,08	5.430	98,76	5.902	98,01
Bókmenntatextar frá því eftir 1980						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	280	7,21	3.601	92,79	3.881	100,00
MXP	182	0,00	3.217	89,34	3.399	87,58
fnTBL	157	56,07	3.262	90,59	3.419	88,10
TnT	221	78,93	3.385	94,00	3.606	92,91
TnT, einf.	221	0,00	3.385	94,00	3.606	92,91
MXP, orðfl.	236	84,29	3.512	97,53	3.748	96,57
fnTBL, orðfl	221	78,93	3.537	98,22	3.758	96,83
TnT, orðfl.	257	91,79	3.561	98,89	3.818	98,38
Textar um tölvur og tækni						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	442	15,11	2484	84,89	2.926	100,00
MXP	186	42,08	2191	88,20	2.377	81,24
fnTBL	169	38,24	2190	88,16	2.359	80,62
TnT	222	50,23	2317	93,28	2.539	86,77
TnT einf.	222	50,23	2317	93,28	2.539	86,77
MXP, orðfl	364	82,35	2410	97,02	2.774	94,81
fnTBL, orðfl	356	80,54	2437	98,11	2.793	95,45
TnT, orðfl.	395	89,37	2453	98,75	2.848	97,33
Textar um lögfræði og viðskipti						
	Óþekkt orð		Þekkt orð		Öll orð	
Aðferð	Tíðni	%	Tíðni	%	Tíðni	%
Alls	390	14,05	2.386	85,95	2.776	100,00
MXP	236	60,51	2.042	85,58	2.278	82,06
fnTBL	213	54,62	2.041	85,54	2.254	81,20
TnT	284	72,82	2.174	91,11	2.458	88,54
TnT einf.	284	72,82	2.176	91,20	2.460	88,62
MXP, orðfl.	348	89,23	2.301	96,44	2.649	95,43
fnTBL, orðfl	336	86,15	2.309	96,77	2.645	95,28
TnT, orðfl.	366	93,85	2.337	97,95	2.703	97,37

Í töflu 2 sjást helstu niðurstöður mörkunar lesmálsorða í þessum textum. Hér kemur í ljós að TnT-markarinn nær bestum árangri. Markararnir MXPOST og fnTBL ná svo lélegum árangri að ekki reyndist unnt að bæta niðurstöðu TnT-markarans með því að nýta niðurstöður frá hinum mörkurunum tveimur. TnT-markarinn nær betri árangri við mörkun bókmenntatextanna heldur en við mörkun texta orðtíðnibókarinnar sjálfrar en verri árangri við mörkun textanna um tölvur og tækni og viðskipti og lögfræði.

Ekki var notað viðbótarorðasafn þannig að óþekkt orð eru þau orð sem ekki koma fyrir í textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hátt í öllum textunum og hærra en meðalhlutfall í prófunarsöfnum sem gerð voru úr textum Orðtíðnibókarinnar. Hlutfall óþekktra orða er hæst í textanum um tölvur og tækni og þar er árangur mörkunar slakastur. TnT-markarinn nær samt alls staðar viðunandi árangri ef aðeins er gerð krafa um réttan orðflokk.

Þessar niðurstöður benda til þess að nauðsynlegt sé að bæta árangur mörkunar óþekktra orða til þess ná viðunandi árangri í mörkun texta. Ein leið til þess að gera það er að hafa til umráða umfangsmiklar orðaskrár þar sem fram koma beygingarmyndir sem flestra orða og mörk þeirra. Nota má beygingarlýsingu íslensks nútímamáls, sem einnig var gerð fyrir styrk frá tungutækniverkefni menntamálaráðuneytisins, sem efnivið í slíka orðaskrá. Einnig er nauðsynlegt að hafa tiltækar skrár með ýmiss konar sérnöfnum svo sem mannanöfnum, nöfnum fyrirtækja og stofnana og örnefnum.

Umsjón með verkinu

Verktakar við verkið voru Málgreiningarhópurinn (Auður Þórunn Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Sigrún Helgadóttir) og Orðabók Háskólans. Gerður var samningur milli ráðuneytisins og verktakanna 18. október 2002. Verkið hófst haustið 2002 og lokaskýrslu var skilað til menntamálaráðuneytisins í febrúar 2004. Verkefnisstjóri var Eiríkur Rögnvaldsson en Sigrún Helgadóttir mótaði vinnulag við prófun markaranna og vann meginhluta vinnunnar ásamt félögum í Málgreiningarhópnum. Orðabók Háskólans lagði til verksins aðstöðu og markað textasafn *Íslenskrar orðtíðnibókar*.

Heimildir

Jörgen Pind (ritstj.), Friðrik Magnússon, Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

Viðauki

Skýring skammstafana í greiningarstrengjum Orðtíðnibókar

Dálkur	Formdeild	Greiningartákn-greiningaratriði
1	Orðflokkur	N-nafnorð
2	kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn, X-ókyngreint
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
5	Greinir	G-með viðskeyttum greini
6	Sérnöfn	M-mannsnafn, Ö-örnefni, S-önnur sérnöfn
1	Orðflokkur	L-lýsingarorð
2	Stig	F-frumstig, M-miðstig, E-efstastig
3	Beyging	S-sterk beyging, V-veik beyging, O-óbeygt
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	Tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	F-fornafn
2	Flokkur	A-ábendingarfornafn, B-óákveðið ábendingarfornafn, E-eignarfornafn O-óákveðið fornafn, P-persónufornafn, S-spurnarfornafn, T-tilvísunarfornafn
3	Kyn/Persóna	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	G-greinir
2	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
3	Tala	E-eintala, F-fleirtala
4	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	T-töluorð
2	Flokkur	F-frumtala
3	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn/1-1. pers., 2-2. pers.
4	Tala	E-eintala, F-fleirtala
5	Fall	N-nefnifall, O-þolfall, Þ-þágufall, E-eignarfall
1	Orðflokkur	S-sögn (þó ekki lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	N-nafnh., B-boðh., F-framsöguh., V-viðtengingarh., S-sagnbót, L-lýsingarh. nútíðar
4	Tíð	N-nútíð, Þ-þátíð
5	Tala	E-eintala, F-fleirtala
6	Persóna	1-1. persóna, 2-2. persóna, 3-3. persóna
1	Orðflokkur	S-sögn (lýsingarháttur þátíðar)
2	Mynd	G-germynd, M-miðmynd
3	Háttur	Þ-lýsingarháttur þátíðar
4	Kyn	K-karlkyn, V-kvenkyn, H-hvorugkyn
5	tala	E-eintala, F-fleirtala
6	Fall	N-nefnifall, O-þolfall
1	Orðflokkur	A-atviksorð
2	Stig	M-miðstig, E-efsta stig
3	Flokkur/Fallstjórn	A-stýrir ekki falli, U-upphrópun/ O-stýrir þolfalli, Þ-stýrir þágufalli E-stýrir eignarfalli
1	Orðflokkur	C-samtenging
2	Flokkur	N-nafnháttarmerki, T-tilvísunartenging
1	Flokkur	E-erlent orð
1		X-ógreint orð