# Formáli

# 1 Inngangur

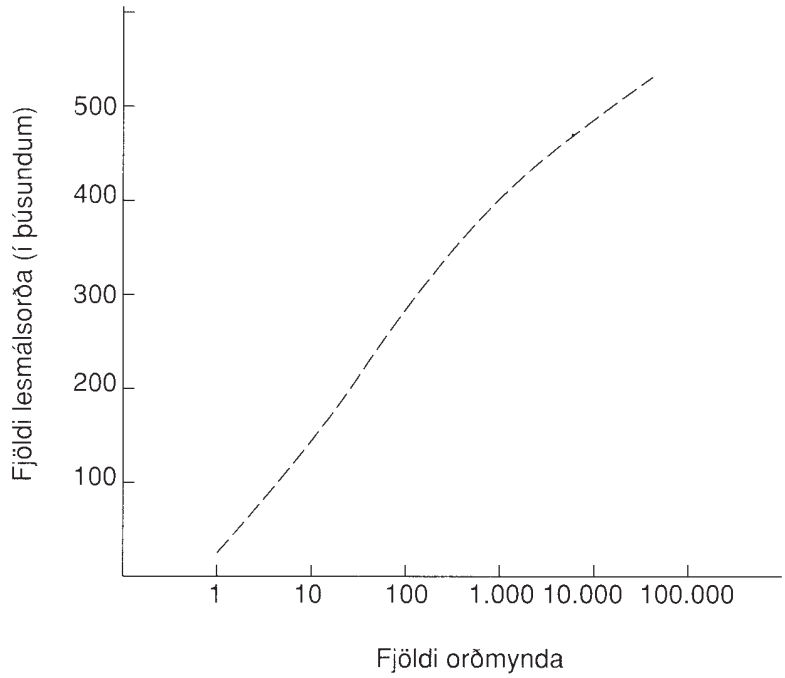
## 1.1 Orðtíðni

Fræðimenn hafa löngum og af margvíslegum ástæðum reynt að gera sér grein fyrir orðaforða einstakra tungumála. Stundum hefur orðtíðni orðið þeim sérstakt áhugaefni. Það eru gömul sannindi að orð eru misjafnlega algeng í tungumálum, sum koma fyrir með stuttu millibili í ræðu og riti, eftir öðrum getur þurft að leita stundarlengi vilji menn finna dæmi um notkun þeirra. Þessi sannindi eru orðabókarmönnum vel kunn, um sum orð er einkar auðvelt að fræðast en erfitt getur verið að öðlast haldgóða vitneskju um önnur. Þetta kemur vel fram á mynd I á næstu síðu en þar er sýnt hvernig háttað er hlutfalli milli orðmynda og lesmálsorða í þeirri orðtíðnikönnun sem greint er frá í þessari bók. Alls voru teknir til athugunar textar með rúmlega 500.000 lesmálsorðum. Af línuritinu kemur fram að 100 algengustu orðmyndirnar koma fyrir tæplega 300.000 sinnum. Slíkt samband er einnig vel þekkt úr öðrum könnunum.

Ekki er þó ýkja langt síðan hafist var handa við eiginlegar rannsóknir á tíðni orða. Athyglisvert er að áhugi á þeim vaknaði um svipað leyti hjá tveim stéttum manna, hraðriturum og uppeldisfræðingum. Fyrstu kannanir munu hafa verið gerðar af hraðriturum sem áttuðu sig á mikilvægi þess að hafa glögga vitneskju um þau orð sem eru algengust í málinu. Þekktar eru hinar umfangsmiklu kannanir Þjóðverjans F. W. Kaedings, sem taldi tíðni orða í rúmlega tíu milljón orða safni árið 1898. Meðal uppeldisfræðinga vaknaði áhugi á orðtíðnirannsóknum í tengslum við móðurmálskennslu barna. Kunnar eru rannsóknir bandaríska uppeldisfrömuðarins E. Thorndikes á tíðni orða í bandarískri ensku sem birtust fyrst árið 1921. Af síðari rannsóknum Thorndikes hefur sú sem birtist í bók hans og Lorges, *The Teacher's Word Book of 30,000 Words*, sem út kom árið 1944, sennilega náð mestri útbreiðslu. Athyglisvert er að Thorndike lagði síðar rannsóknir sínar á orðtíðni til grundvallar umfangsmikilli orðabókagerð í Bandaríkjunum (Landau 1989).[1]

Fregnir af þessum rannsóknum Thorndikes og annarra bárust hingað til lands og urðu kveikja að íslenskum orðtíðnirannsóknum sem Ársæll Sigurðsson skólastjóri stóð að. Þegar árið 1938 ritaði hann grein í *Menntamál: Tímarit um uppeldis- og fræðslumál* þar sem hann lýsir þörf á því að rannsaka íslenska orðtíðni vegna móðurmálskennslunnar, sérstaklega til þess að komast að því hver séu algengustu orð málsins eða „frumorðin".

1 Hin ágæta ensk-íslenska orðabók Arnar og Örlygs á rætur að rekja til einnar af orðabókum Thorndikes.




*Mynd I. Dæmigert samband orðmynda og lesmálsorða. Hundrað algengustu orðmyndirnar eru meira en helmingur ritaðra orða í venjulegum textum.*

Honum farast m.a. svo orð (Ársæll Sigurðsson 1938:101):

> Það virðist augljóst, hversu mikill stuðningur það væri fyrir kennarann að hafa þessi „frumorð“ málsins til hliðsjónar við kennsluna, þótt ekki væri fleiri en 500–1000. Þá væri það einnig mikilsvert atriði, ef hægt væri að komast að raun um, hvaða orð börnum er tamast að nota, þegar þau gera grein fyrir einhverju efni skriflega eða munnlega, en það er þróunarsaga hins lifandi máls hjá hverjum einstaklingi, á hvern hátt sú tjáning fer fram á ýmsum aldursskeiðum.

Í sömu grein getur Ársæll þess svo að hann sé þegar byrjaður á slíkri könnun. Niðurstöður hennar birtust síðan í *Menntamálum* árið 1940. Könnun Ársæls var um margt gagnmerk og verður nánar að henni vikið hér á eftir.

En áhugi á orðtíðni nær einnig til fleiri fræðasviða en uppeldisfræðinnar. Fyrr er minnst á áhuga hraðritara á orðtíðni. Sama áhuga verður vart hjá upplýsinga- og tölvufræðingum en þeir hafa mjög kannað tölfræðilegar eigindir tungumála, m.a. í því skyni að auðvelda mat á því með hvaða hætti megi þjappa rituðu máli saman í tölvugeymslu og hvernig hægt sé að útbúa dullykla sem gera mönnum kleift að geyma ritmál á öruggan hátt. Á seinustu árum hefur einnig verið þörf fyrir vitneskju um orðtíðni við gerð hugbúnaðar á borð við villuleitarforrit (McIlroy 1982).

Málfræðingar hafa kannski ekki sinnt orðtíðnirannsóknum í þeim mæli sem búast hefði mátt við en þó hefur mátt greina aukinn áhuga meðal þeirra á slíkum rannsóknum á seinni árum. Hér skiptir vafalítið töluverðu máli að aðgangur fræðimanna að tölvutæku efni hefur stórbatnað eftir að ritvinnsla á tölvum komst á fullan skrið. Er nú svo komið að næsta auðvelt er fyrir nánast hvern sem er að koma sér upp dágóðu safni texta í tölvutækri mynd.

Með tilkomu tölvutækninnar hefur orðið auðveldara en áður að stunda rannsóknir á orðtíðni. Því fer þó fjarri að tölvan létti öllu erfiði af fræðimönnum eins og væntanlega mun koma glöggt fram hér síðar þegar lýst verður framkvæmd þessarar könnunar. Fyrstur til að rannsaka orðtíðni hér á landi með aðstoð tölvu var Baldur Jónsson prófessor sem kannaði tíðni orða í skáldsögu Ólafs Jóhanns Sigurðssonar *Hreiðrinu* á árunum 1973–1974. Síðan hafa allmargar slíkar athuganir verið gerðar og verður hinna helstu þeirra getið hér á eftir auk þess sem nákvæmlega verður greint frá framkvæmd og tilhögun þeirrar könnunar sem liggur til grundvallar *Íslenskri orðtíðnibók*.

Aðdraganda þeirrar könnunar sem birtist á þessari bók má rekja allt til ársins 1984 þegar Orðabók Háskólans tók að sér að útbúa orðasafn fyrir villuleitarforrit sem IBM á Íslandi hugðist koma á markað hér á landi. Eins og áður er getið er nauðsynlegt að taka fullt tillit til orðtíðni við gerð slíks orðasafns. Sú staðreynd að 100 orðmyndir eða svo eru um helmingur allra lesmálsorða í venjulegum rituðum texta hefur áhrif á það með hvaða hætti gera á slíkt orðasafn úr garði svo að villuleitin verði sem hraðvirkust. Einn þáttur í gerð villuleitarforritsins var sérstök orðtíðnikönnun. Sumir textanna sem þá voru kannaðir eru einnig notaðir í þessari bók.

Skömmu síðar var ákveðið að Orðabókin kæmi sér upp safni af tölvutækum textum til að nota við gerð hinnar sögulegu orðabókar sem unnið er að á stofnuninni. Hér höfðu menn ekki síst í huga notagildi slíks textasafns við að leita dæma um orðanotkun. Textasafnið hefur vaxið mjög á undanförnum árum og reynst með þeim hætti sem vænst var. Í framhaldi af þessu var ákveðið að ráðast í vandaða orðtíðnikönnun þar sem þess yrði freistað að gera tíðni orða í nútímamáli rækileg skil. Samþykkti stjórn Orðabókarinnar í apríl 1985 að ráðast í slíkt verk. Fyrir stjórninni vakti fyrst og fremst að með þessu móti yrði lagt nokkuð af mörkum til rannsókna á íslensku nútímamáli.

## 1.2 Fyrri rannsóknir

### 1.2.1 Erlendar rannsóknir

Eins og áður hefur verið greint frá var það bandaríski uppeldisfræðinguri nn Thorndike sem einna fyrstur varð til þess að rannsaka orðtíðni. Ekki er ætlunin að gera hér neina viðhlítandi grein fyrir erlendum rannsóknum á tíðni orða í ólíkum málum en þó er rétt að víkja að tveim erlendum rannsóknum sem höfð var nokkur hliðsjón af við gerð þessarar könnunar.

Á 7. áratugnum var hafist handa við merka orðtíðnirannsókn við Brownháskólann í Bandaríkjunum sem síðan hefur verið kennd við þann skóla. Þessi rannsókn er ekki hvað síst athyglisverð fyrir fernt:

1. Textar í könnuninni voru valdir af mismunandi efnissviðum.
2. Beitt var tölfræðilegum aðferðum við val á texta innan hvers efnisflokks.
3. Öll orðin í textunum hlutu nákvæma málfræðilega greiningu.
4. Rannsakendur við Brown-háskólann hafa heimilað öðrum aðilum afnot af safni sínu („Brown Corpus“) og því hafa fræðimenn getað notað þetta safn við margvíslegar athuganir.

Alls tók Brown-rannsóknin til 15 efnisflokka, m.a. frétta, trúarbragða, skáldskapar, vísinda og gamanmála. Í hverjum efnisflokki voru nokkur textasýni og var hvert þeirra um 2.000 lesmálsorð að lengd. Brown-safnið er alls 1 milljón lesmálsorða og því eru samtals 500 textasýni í því. Vægi einstakra flokka er hins vegar mismunandi.

Öll textasýnin voru frá árinu 1961 en eftir að búið var að ákveða vægi efnisflokka voru textasýnin fundin með slembiúrtaki.

Sjálf málfræðigreiningin fór fram nokkru síðar eða á árunum 1970–1978. Niðurstöður rannsóknarinnar birtust í endanlegri gerð árið 1982 (Francis og Kučera 1982).

Brown-rannsóknin tók til bandarískrar ensku. Árið 1970 var byrjað á hliðstæðu verki við háskólann í Lancaster í Bretlandi með það að markmiði að gera svipaða könnun á breskri ensku. Sú rannsókn stóð lengi og lauk henni í Noregi í höndum þeirra Stig Johanssons og Knut Hoflands. Birtu þeir helstu niðurstöður á bók árið 1989 (Johansson og Hofland 1989a, 1989b). Þessi rannsókn tók einnig til milljón lesmálsorða í 500 textasýnum og 15 efnisflokkum enda var henni ætlað að vera nothæf til samanburðar á breskri og bandarískri ensku. Í kjölfar hennar hafa síðan verið birtar athyglisverðar rannsóknir á vélrænni greiningu texta. Brown-safnið var greint með vélrænum hætti á sínum tíma en árangurinn var ekki mjög góður. Langtum betri árangur náðist hins vegar í greiningu á breska safninu eins og skýrt er frá hjá Garside, Leech og Sampson (1987).

Aðferðafræði þessara tveggja rannsókna hefur með beinum og óbeinum hætti sett mark sitt á margar aðrar tíðnikannanir. Hér má t.d. nefna rann-

sóknir Sture Alléns á tíðni orða í sænskum dagblöðum (Allén 1970, 1971) og einnig rannsókn Gunnel Engwalls á tíðni orða í frönskum skáldsögum áranna 1962–1968 (Engwall 1984).

Allar þær rannsóknir sem hér hafa verið nefndar voru nokkur hvati fyrir þá könnun sem hér er greint frá. Þannig einsettum við okkur í upphafi að velja texta úr ólíkum efnisflokkum (nánar er greint frá þeim síðar) sem og að greina beygingu allra orða nákvæmlega.

Nánar er greint frá aðferðafræði þessarar könnunar hér á eftir.

### 1.2.2 Íslenskar rannsóknir

#### 1.2.2.1 Rannsókn Ársæls Sigurðssonar

Það mun hafa verið Ársæll Sigurðsson skólastjóri sem fyrstur manna rannsakaði orðtíðni í íslensku nútímamáli og birti hann niðurstöður sínar í greininni *Algengustu orðmyndir málsins og stafsetningarkennslan* í *Menntamálum* árið 1940. Í þessari rannsókn leitaðist Ársæll við að ákvarða algengasta orðaforða málsins „í þeim höfuðtilgangi að finna leið til að gera stafsetningarkennsluna aðgengilegri og raunhæfari en áður, en þó vænlegri til betri árangurs“ (Ársæll Sigurðsson 1940:9–10). Efnivið sinn valdi Ársæll með hliðsjón af þessu markmiði úr eftirtöldum efnisflokkum: stílum barna, sendibréfum fullorðinna, lesbókum, náttúrufræði, sögu og landafræði.

Helstu niðurstöður úr rannsókn Ársæls Sigurðssonar eru sem hér segir. Alls reyndust lesmálsorðin vera 100.227, þar af 3.530 eiginnöfn (3,52%) og 49 ártöl (0,05%). Mismunandi orðmyndir voru 13.636.

Könnunin leiddi m.a. í ljós að 100 algengustu orðmyndirnar, sem eru aðeins 0,73% af orðmyndafjöldanum, koma fyrir 50.263 sinnum sem er 52,01% alls lesmálsins. Einnig kemur fram að meirihluti orðmyndanna, eða 60,48%, kemur aðeins fyrir einu sinni.

Um þessar niðurstöður segir Ársæll Sigurðsson (1940:19–20):

> Samanburður þessi sýnir ljóslega, að það eru tiltölulega fáar orðmyndir, sem mynda meginhluta lesmálsins, en aftur á móti verður meiri hluti einstakra orðmynda aðeins lítill hluti af lesmálinu. […]
>
> Þetta sannar því það, að um íslenzkuna gildir hið sama og um aðrar menningartungur, að ritað mál (talmál sennilega ekki síður) er samsett af fáum orðmyndum, sem oft eru endurteknar, og mörgum orðmyndum, sem sjaldan eða aldrei eru endurteknar.

#### 1.2.2.2 Rannsókn Baldurs Jónssonar

Rannsókn Ársæls Sigurðssonar sem sagt var frá hér að framan var, eins og nærri má geta, unnin „í höndunum“ þar sem engar tölvur var að fá til aðstoðar á þeim tíma. Tölvur voru hins vegar komnar fram á sjónarsviðið í upphafi áttunda áratugarins þegar Baldur Jónsson, Björn Ellertsson og

Sven Þ. Sigurðsson fóru af stað með könnun sína á tíðni orða í skáldsögu Ólafs Jóhanns Sigurðssonar *Hreiðrinu*. Vinna við hana fór einkum fram á árunum 1973–74 og birtust niðurstöður hennar síðan í þrem tíðniskrám árið 1975, *Tíðni orða í Hreiðrinu 1–3*, og síðan fimm árum síðar í greinargerð um könnunina með frekari niðurstöðum undir heitinu *Tölvukönnun á tíðni orða og stafa í íslenskum texta*. Annar afrakstur þessarar vinnu var *Orðstöðulykill að Hreiðrinu* sem út kom árið 1978.

Skipta má tíðnikönnuninni í tvennt: Annars vegar könnun á tíðni orðmynda en hins vegar könnun á tíðni rittákna, sambandi orðlengdar og orðtíðni, meðallengd orðmynda, meðaltíðni orðmynda o.fl. slíkt. Niðurstöður úr orðtíðnikönnuninni birtust í áðurnefndum þrem tíðniskrám (Baldur Jónsson 1975). Sú fyrsta hefur að geyma orðmyndir í stafrófsröð eftir upphafi þeirra, önnur orðmyndir í stafrófsröð eftir niðurlagi þeirra og hin þriðja orðmyndir í röð eftir lækkandi tíðni. Tölulegar niðurstöður birtust síðan í greinargerðinni (Baldur Jónsson o.fl. 1980:55–126) í skrám og töflum. Meðal upplýsinga sem þar er að finna er samanburður á tíðni 100 algengustu orðmynda í *Hreiðrinu* og orðasafni Ársæls Sigurðssonar, tíðni bókstafa, skipan sérhljóða og samhljóða í algengustu gerðum orðmynda og lesmálsorða, tíðni tvístöfunga svo og heildaryfirlit yfir tíðni orðmynda og rittákna.

Lesmálsorð reyndust vera 53.226 og orðmyndir 11.341 þannig að meðaltíðni orðmynda var 4,70.

#### 1.2.2.3 Aðrar íslenskar athuganir

Ekki er ástæða til að fjalla rækilega um aðrar tíðnikannanir íslenskar (nánar er um sumar þeirra fjallað í grein eftir Friðrik Magnússon í tímaritinu *Orð og tunga*, 1. árgangi) en þó er rétt að nefna nokkrar hér sem fjallað hefur verið um á prenti.

Árið 1979 birtu Indriði Gíslason og Sigríður Valgeirsdóttir niðurstöður úr könnun á tíðni viðtengingarháttar í þátíð í fyrsta árgangi tímaritsins *Íslenskt mál*. Eins og Ársæll Sigurðsson höfðu þau Indriði og Sigríður kennslufræðileg sjónarmið að leiðarljósi við gerð könnunarinnar (Indriði Gíslason og Sigríður Valgeirsdóttir 1979:107–8):

> Kennurum og verðandi kennurum þótti forvitnilegt að kanna tíðni viðtengingarháttar í þátíð en vitað er að í móðurmálskennslu fer umtalsverður tími til að kenna rithátt slíkra sagnmynda. Verður nemendum einkum villugjarnt í vth. þt. af sterkum sögnum en sé tekið mið af kennslubókum í stafsetningu og samræmdum prófum í þeirri grein virðist miklu púðri eytt á ýmsar orðmyndir af þessu tagi án tillits til tíðni þeirra.

Efniviður könnunarinnar var annars vegar 40 tölublöð fjögurra dagblaða frá árinu 1975, alls 49.371,7 dálksentimetrar, og hins vegar 16 tölublöð fjögurra dagblaða frá árinu 1925, alls 7.673,8 dálksentimetrar. Reiknaður

fjöldi lesmálsorða út frá fjölda dálksentimetra er 738.106,9 frá árinu 1975 og 114.723,4 frá árinu 1925.

Helstu niðurstöður eru þær að sagnirnar *vera, hafa* og *verða* koma oftast fyrir í viðtengingarhætti í þátíð og eru þær samanlagt 57,6% allra sagna í viðtengingarhætti í þátíð árið 1925 og 63% árið 1975. Þá draga þau Indriði og Sigríður þær ályktanir af niðurstöðum sínum að ekki verði séð að notkun viðtengingarháttar í þátíð fari dvínandi.

Loks má geta þess að á síðustu árum hefur nokkuð verið unnið við rannsóknir á orðaforða Eddukvæða (Baldur Jónsson 1990), Biblíunnar (Baldur Pálsson 1990) og Íslendingasagna (Eiríkur Rögnvaldsson 1990).

# 2 Hugtök

Í fyrri kafla brá fyrir nokkrum hugtökum sem hafa e.t.v. verkað torkennilega á lesandann. Rætt var um lesmálsorð og orðmyndir, orð og beygingarmyndir án þess að því fylgdu nánari skýringar. Er rétt að freista þess að skilgreina nokkur þau hugtök sem helst koma við sögu í orðtíðnirannsóknum.

Hugtakið **orð** er býsna margrætt og gera verður greinarmun á margvíslegri merkingu sem það getur falið í sér. Lesandinn kannast vafalítið við að ritsíminn hefur tamið sér að telja fjölda orða í símskeytum. Oftast mundi ekki vefjast fyrir mönnum að segja til um þennan fjölda en þó getur það nokkuð oltið á rithætti hvernig á að telja. Oft eru menn í vafa um hvort rita eigi í einu orði eða tveim ýmsar samtengingar og algeng smáorð. Þannig er algengt að menn riti *alltof* í stað *allt of*, *einskonar* í stað *eins konar* o.s.frv. Oft eru einnig áhöld um hvernig rita skuli nafnorð. Þannig er orðið *Íslendingasögur* oftast ritað sem eitt orð, en þó er einnig til að það sé ritað í tveim orðum: *Íslendinga sögur*. Enn vandast þó málið þegar hugað er að orðum sem eru sett saman úr öðru en bókstöfum úr stafrófinu. Telst $H_2O$ t.d. vera orð? Og hvað með stafastrengi á borð við $(d/2c)(mc^2/E)^2$, *°C* eða $km^2$? Ekki er til nein einföld og einhlít aðferð til að draga mörk milli þess sem hægt er réttu lagi að kalla orð og hins sem eru einfaldlega stafarunur sem hafa meira eða minna óljós tengsl við orðaforða málsins.

Þegar svo háttar til að menn koma sér ekki saman um hvar draga eigi mörk orða getur verið erfitt að segja til um hversu mörg orð eru í tilteknum texta. Hér er þó einvörðungu verið að fjalla um orðin eins og þau eru skrifuð á blað, ritorðin. Málið vandast enn frekar þegar hugað er að því að flokka orðin. Þá er spurt sem svo hvort eitt ritað orð sé „sama“ orðið og annað orð. Eru orðin *er* og *var* t.d. sama orðið? Já, í vissum skilningi eru þau það. Í báðum tilvikum er um að ræða beygingarafbrigði sama orðs, þ.e. sagnarinnar *vera*. Í öðru tilliti er augljóslega ekki um að ræða „sama“ orðið. Annað er skrifað *e-r* en hitt *v-a-r*. Hér er því þörf á því að skoða hvaða merkingar er hægt að leggja í orðið **orð**.

Hér kemur einnig annað til. Í stafarunu á borð við *BSRB* eru fjórir stafir en þó einvörðungu þrír mismunandi stafir. Gera verður skýran greinar-

mun á þeim tilvikum þar sem telja á stafi, orð eða önnur fyrirbæri án tillits til endurtekningar og þeim tilvikum þar sem einungis er ætlunin að telja hversu mörg mismunandi fyrirbæri eru á ferðinni. Í ýmsum erlendum málum eru hugtökin **type** og **token** notuð um þessa aðgreiningu. Fyrra orðið hefur verið þýtt sem tegund eða tag á íslensku, en hið seinna sem tákn eða stak (og jafnvel með hvorugkynsnafnorðinu tók). Í stafastrengnum *BSRB* eru þá fjögur tákn eða stök en aðeins þrjú tög. Sama aðgreining skiptir máli þegar kemur að því að skoða orð í setningum.

Ef skoðuð er stutt setning á borð við þessa:

> Ég minni ykkur á það sem málfræðingurinn sagði í áheyrn minni: Gætið að orðunum, málfræðingar!

liggur beint við að segja sem svo að þar fari 15 orð. Ef orðtalningarforrit er látið telja orðin í þessari setningu fáum við nákvæmlega þá tölu: 15. Hér er hugtakið **orð** notað í sama skilningi og hugtakið stak. Í íslensku er venjan að kalla slík orð **lesmálsorð**. Þá er sagt sem svo að í fyrrgreindri setningu séu lesmálsorðin 15. Í þeirri orðtíðnikönnun sem hér er sagt frá reyndust lesmálsorðin alls vera 519.186. Til samanburðar má geta þess að Njáls saga er tæplega 100.000 lesmálsorð.

Aðgreining eins orðs frá öðru er byggð á rithætti orðanna og er tíðast fylgt reglum á borð við þá að telja samfellda táknstrengi milli greinarmerkja til sérstakra orða. Nánar er greint frá þeim reglum sem fylgt var í þessari könnun við afmörkun lesmálsorða í grein 3.1.4 í formála. Hugtakið lesmálsorð er stakhugtak. En orðin raða sér einnig í tegundir eða tög og geta reyndar gert það með mismunandi hætti eftir því hvernig orðin eru skoðuð.

Í fyrsta lagi er hugsanlegt að skoða aðeins ritmyndir orðanna óháð merkingu þeirra eða beygingu. Sé það gert kemur í ljós að í fyrrgreindri setningu eru 14 ólík orð í þeim skilningi því eitt orðanna, *minni*, kemur tvisvar fyrir. Hugtakið **orðmynd** er notað um orð í þessum skilningi og samkvæmt því eru 14 ólíkar orðmyndir í þessari setningu. Flestar orðmyndanna í fyrrgreindri setningu koma aðeins fyrir einu sinni, en ein þó tvisvar.

Tölvum lætur einkar vel að telja bæði lesmálsorð og orðmyndir í tölvutækum textum og því er það að mjög margar tíðnirannsóknir, sem unnar hafa verið í tölvu, takmarkast við það að telja þessi tvö fyrirbæri. Við það er þó litið fram hjá mikilsverðum upplýsingum. Ef fyrrgreind setning er skoðuð nánar sjáum við t.d. að þar er að finna tvö beygingarafbrigði orðsins *málfræðingur*. Annað er *málfræðingurinn* sem stendur í nefnifalli eintölu með greini, hitt er *málfræðingar* sem stendur í nefnifalli fleirtölu og er án greinis. Full ástæða væri til að flokka þau orð saman. Einnig sést að orðmyndaflokkunin fellir saman tvær orðmyndir, þar sem ekki er um sama orðið að ræða í skilningi orðabóka. Er hér átt við orðmyndina *minni*. Í fyrra tilvikinu, „ég minni“, er um sögn að ræða en í hinu síðara, „áheyrn minni“, er um að ræða fornafn. Hin vélræna flokkun orðanna gerir hér engan greinarmun á enda byggir hún einvörðungu á rithætti orðmyndanna.

Hugtakið **flettiorð** er hér notað um orð sem eru ólík að málfræðilegri flokkun. Hér er því um að ræða annað tegundarhugtak fyrir orð sem er ólíkt hugtakinu **orðmynd** þar sem flokkunin byggist einvörðungu á rithætti orðanna. Greiningin í flettiorð byggist á málfræðilegri flokkun orðanna. Þannig eru orð af ólíkum orðflokkum flokkuð sem aðgreind flettiorð. Ef um nafnorð er að ræða skiptir kyn þess einnig máli. Þannig er nafnorðið *hlið* bæði til í kvenkyni og hvorugkyni og teljast þau því tvö ólík flettiorð þrátt fyrir að útlit orðanna sé hið sama, a.m.k. í sumum beygingarmyndum.

Flokkun lesmálsorða í flettiorð er ólíkt flóknari en flokkun þeirra í orðmyndir. Tölvur geta flokkað lesmálsorð í orðmyndir án fyrirhafnar enda byggir sú flokkun einvörðungu á samanburði á rithætti. Flokkun í flettiorð þarf að byggja á verulegri málfræðilegri þekkingu og slík þekking er tölvum ekki sjálfkrafa gefin. Könnun sú sem greint er frá í þessu riti markar tímamót í íslenskum orðtíðnirannsóknum að því leyti að hér er orðaforði rannsóknarinnar í fyrsta skipti greindur í flettiorð. Sú greining fór að hluta til fram með vélrænum hætti en að hluta með „handvirkum“ aðferðum eins og nánar er greint frá aftar.

Að síðustu er rétt að skilgreina einnig hugtakið **greiningarmynd** sem er töluvert notað í þessari rannsókn. Með greiningarmynd er átt við orðmynd ásamt málfræðilegum „greiningarstreng“. Í aðalkafla bókarinnar (bls. 3–554) er að finna allar orðmyndir flokkaðar eftir flettiorðum en auk þess eru orðmyndirnar aðgreindar ef þær hafa ólíka beygingu þrátt fyrir að ritmynd þeirra sé hin sama. Svo dæmi sé tekið er orðmyndin *haust* færð þrisvar undir flettiorðið *haust* (bls. 182). Fyrsta dæmið er um orðið í nefnifalli eintölu, annað dæmið er um orðið í þolfalli eintölu og loks kemur það líka fyrir í þolfalli fleirtölu. Allar þessar beygingarmyndir eru táknaðar með sömu orðmyndinni, *haust*. Ef einvörðungu hefði verið ætlunin að greina flettiorð í þessari könnun hefði mátt sameina allar þessar beygingarmyndir í eina undir flettiorðinu. En þar eð ætlunin var að ná einnig til allra beygingarmynda orðsins er þeim haldið aðgreindum. Við notum hugtakið **greiningarmynd** um orðmynd ásamt hinum málfræðilega greiningarstreng.

Svo enn sé vikið að setningunni

> Ég minni ykkur á það sem málfræðingurinn sagði í áheyrn minni:
> Gætið að orðunum, málfræðingar!

þá kemur í ljós að **lesmálsorð** í henni eru alls 15, **orðmyndir** 14, **flettiorð** 14 og **greiningarmyndir** 15.

# 3 Orðtíðnikönnun Orðabókar Háskólans

Eins og fyrr er greint frá er nú nokkuð um liðið síðan orðtíðnikönnun Orðabókar Háskólans hófst. Í þessum kafla er nánar greint frá framkvæmd könnunarinnar og efniviði hennar.

## 3.1 Framkvæmd könnunarinnar

### 3.1.1 Textaval

Könnunin nær til 100 texta sem eru svipaðir að stærð, u.þ.b. 5.000 lesmálsorð hver. Í langflestum tilvikum eru textarnir hluti af stærra ritverki.

Textar voru valdir úr ritverkum sem gefin voru út á áratugnum 1980–1989. Ef notuð var önnur útgáfa en hin fyrsta var þess gætt að fyrsta útgáfa textans væri einnig frá þessu árabili. Þegar um er að ræða texta sem þýddur er úr erlendu máli er miðað við útgáfuár þýðingarinnar en ekki frumtextans.

Í upphafi var ákveðið að velja texta úr fimm textaflokkum, jafnmarga texta úr hverjum flokki eða tuttugu alls. Flokkarnir voru eftirfarandi:

1. Íslensk skáldverk.
2. Þýdd skáldverk.
3. Ævisögur og endurminningar.
4. Fræðslutextar.
5. Barna- og unglingabækur.

Textar í 4. textaflokki skiptast jafnt milli fræðsluefnis á sviði hugvísinda (tíu textar) og fræðsluefnis á sviði raunvísinda og tækni (tíu textar). Textar í 5. textaflokki skiptast jafnt milli frumsamins barnaefnis (tíu textar) og þýdds barnaefnis (tíu textar).

Þau skilyrði voru sett að höfundur(-ar) og þýðandi(-endur) hvers texta væru hvorki höfundar né þýðendur að öðrum textum í könnuninni.

Að öðru leyti var val textanna látið ráðast af því hversu auðfengin ritin væru og þá einkum af því hvort þau væru tiltæk tölvuskráð. Flestir textanna voru sóttir í tölvutækt textasafn Orðabókar Háskólans en nokkra þurfti að tölvuskrá sérstaklega vegna þessa verkefnis.

Hver texti hefst á samfelldu máli en ekki t.d. á fyrirsögn eða kaflanúmeri. Þegar texti er valinn sem hluti af stærra ritverki, eins og langoftast er raunin, er hann að jafnaði, en þó ekki alltaf, tekinn úr upphafi meginmáls og síðan sleppt úr myndefni og öðru mjög sundurlausu efni. Fyrirsögnum og kaflanúmerum inni í textunum er þó ekki sleppt. Texti er ávallt látinn enda á heilli setningu, oftast við lok kafla eða annars skýrt afmarkaðs rithluta.

### 3.1.2 Efniviður

Hér fer á eftir skrá um þau rit sem textasýnin voru sótt til. Blaðsíðutöl eru tilgreind ef sýnið er tekið úr texta sem er hluti af stærra verki og sýna þau þá hvar textann er að finna í viðkomandi verki en ekki er tilgreint nákvæmlega hvaðan sýnið er tekið úr textanum.

## 1. Íslensk skáldverk

1 Guðmundur Andri Thorsson. Mín káta angist. Skáldsaga. Mál og menning. Reykjavík, 1988.

2 Svava Jakobsdóttir. Endurkoma. Smásögur Listahátíðar 1986, bls. 113–131. Almenna bókafélagið. Reykjavík, 1986.

3 Steinunn Jóhannesdóttir. Fagrafold. Smásögur Listahátíðar 1986, bls. 133–153. Almenna bókafélagið. Reykjavík, 1986.

4 Sigurður A. Magnússon. Úr snöru fuglarans. Uppvaxtarsaga. Mál og menning. Reykjavík, 1986.

5 Einar Kárason. Gulleyjan. Skáldsaga. Mál og menning. Reykjavík, 1985.

6 Fríða Á. Sigurðardóttir. Eins og hafið. Vaka-Helgafell. 1986.

7 Álfrún Gunnlaugsdóttir. Hringsól. Skáldsaga. Mál og menning. Reykjavík, 1987.

8 Vilhelm Emilsson. Ísis. Smásögur Listahátíðar 1986, bls. 161–178. Almenna bókafélagið. Reykjavík, 1986.

9 Vigdís Grímsdóttir. Kaldaljós. Svart á hvítu. Reykjavík, 1987.

10 Þórarinn Eldjárn. Margsaga. Gullbringa. Reykjavík, 1985.

11 Ólafur Jóhann Ólafsson. Markaðstorg guðanna. Vaka-Helgafell. Reykjavík, 1988.

12 Einar Már Guðmundsson. Riddarar hringstigans. Almenna bókafélagið. Reykjavík, 1982.

13 Úlfar Þormóðsson. Þrjár sólir svartar. Skáldsaga. Útgefandi: Höfundur. 1988.

14 Hermann Másson [Guðbergur Bergsson]. Froskmaðurinn. Forlagið. Reykjavík, 1985.

15 Guðlaugur Arason. Sóla, Sóla. Skáldsaga. Mál og menning. Reykjavík, 1985.

16 Pétur Gunnarsson. Sagan öll. Skáldsaga. Punktar. Reykjavík, 1985.

17 Stefanía Þorgrímsdóttir. Nótt í lífi Klöru Sig. Forlagið. Reykjavík, 1985.

18 Ólafur Gunnarsson. Heilagur Andi og englar vítis. Forlagið. Reykjavík, 1986.

19 Indriði G. Þorsteinsson. Átján sögur úr álfheimum. Almenna bókafélagið. Reykjavík, 1986.

20 Ómar Þ. Halldórsson. Örugglega langur tími. Smásögur Listahátíðar 1986, bls. 185–206. Almenna bókafélagið. Reykjavík, 1986.

## 2. Þýdd skáldverk

21 Bagley, Desmond. Arfurinn. Björn Gíslason íslenskaði. Suðri. Reykjavík, 1982.

22 Morrison, Toni. Ástkær. Úlfur Hjörvar þýddi. Forlagið. Reykjavík, 1988.

23 Morrell, David. Bráð banaráð. Andrés Kristjánsson þýddi. Iðunn. Reykjavík, 1985.

24 Quinnell, A.J. Einfarinn. Spennandi skáldsaga. Björn Jónsson íslenskaði. Almenna bókafélagið. Reykjavík, 1985.

25 Austen, Jane. Hroki og hleypidómar. Silja Aðalsteinsdóttir þýddi. Mál og menning. Reykjavík, 1988.

26 Amado, Jorge. Tvenns konar andlát Kimma vatnsfælna. Sigurður Hjartarson þýddi. Uglan, Íslenski kiljuklúbburinn. Reykjavík, 1989.

27 Fowles, John. Ástkona franska lautinantsins. Magnús Rafnsson þýddi. Mál og menning. Reykjavík, 1985.

28 Allende, Isabel. Eva Luna. Tómas R. Einarsson íslenskaði. Mál og menning. Reykjavík, 1989.

29 MacLean, Alistair. Svarti riddarinn. Sigurður G. Tómasson þýddi. Iðunn. Reykjavík, 1986.

30 Eco, Umberto. Nafn rósarinnar. Thor Vilhjálmsson þýddi. Svart á hvítu. Reykjavík, 1984.

31 Turow, Scott. Uns sekt er sönnuð. Gísli Ragnarson þýddi. Uglan, Íslenski kiljuklúbburinn. Reykjavík, 1989. [2. útg. (1. útg. 1988)]

32 Ludlum, Robert. Svikamyllan. Gissur Ó. Erlingsson þýddi. Setberg. Reykjavík, 1984.

33 Hemingway, Ernest. Gamli maðurinn og hafið. Björn O. Björnsson íslenzkaði. Endurskoðuð þýðing. Almenna bókafélagið. Reykjavík, 1986. [2. útgáfa.]

34 Kemal, Yashar. Memed mjói. Saga um uppreisn og ást. Þórhildur Ólafsdóttir þýddi úr tyrknesku. Mál og menning. Reykjavík, 1985.

35 Weldon, Fay. Ævi og ástir kvendjöfuls. Elísa Björg Þorsteinsdóttir íslenskaði. Forlagið. Reykjavík, 1985.

36 James, P.D. Vitni deyr 1. Álfheiður Kjartansdóttir þýddi. Uglan, Íslenski kiljuklúbburinn. Reykjavík, 1986.

37 Walker, Alice. Purpuraliturinn. Ólöf Eldjárn þýddi. Forlagið. Reykjavík, 1986.

38 Buck, Pearl S. Dætur frú Liang. Arnheiður Sigurðardóttir íslenskaði. Almenna bókafélagið. Reykjavík, 1986.

39 Dostojevskí, Fjodor. Fávitinn. Skáldsaga í fjórum hlutum. Fyrra bindi. Ingibjörg Haraldsdóttir þýddi. Mál og menning. Reykjavík, 1986.

40 Greene, Graham. Tíundi maðurinn. Árni Óskarsson íslenskaði. Almenna bókafélagið. Reykjavík, 1986.

## 3. Ævisögur og endurminningar

41 Gylfi Gröndal. Við byggðum nýjan bæ. Minningar Huldu Jakobsdóttur heiðursborgara Kópavogs. Skráðar eftir frásögn hennar og fleiri heimildum. Almenna bókafélagið. Reykjavík, 1988.

42 Ásgeir Jakobsson. Einar Þorgilsson útgerðarmaður og kaupmaður. Þeir settu svip á öldina. Íslenskir athafnamenn II, bls. 97–111. Ritstjóri Gils Guðmundsson. Iðunn. Reykjavík, 1988.

43 Vilhjálmur Hjálmarsson. Eysteinn í stormi og stillu. Ævisaga Eysteins Jónssonar fyrrum ráðherra og formanns Framsóknarflokksins. III. hluti. Vaka. Reykjavík, 1985.

44 Steinunn Sigurðardóttir. Ein á forsetavakt. Dagar í lífi Vigdísar Finnbogadóttur. Iðunn. Reykjavík, 1988.

45 Ingvi Hrafn Jónsson. Og þá flaug hrafninn. Frjálst framtak. 1988.

46 Stefán Júlíusson. Jóhannes J. Reykdal framkvæmdastjóri og bóndi. Þeir settu svip á öldina. Íslenskir athafnamenn II, bls. 195–209. Ritstjóri Gils Guðmundsson. Iðunn. Reykjavík, 1988.

47 Eðvarð Ingólfsson. Baráttusaga athafnamanns. Endurminningar Skúla Pálssonar á Laxalóni. Æskan. 1988.

48 Sigurjón Björnsson og Aðalsteinn Ingólfsson. Jóhannes Geir. Listasafn ASÍ og Lögberg. Reykjavík, 1985.

49 Þorsteinn Matthíasson. Hrafnistumenn III. Minningar og frásagnir vistmanna og kvenna á Dvalarheimili aldraðra sjómanna, Hrafnistu, skráðar af Þorsteini Matthíassyni. Sjómannadagsráð. 1985.

50 Elín Pálmadóttir. Gerður. Ævisaga myndhöggvara. Almenna bókafélagið. Reykjavík, 1985.

51 Hannes Sigfússon. Framhaldslíf förumanns. Endurminningar Hannesar Sigfússonar skálds. Iðunn. Reykjavík, 1985.

52 Lena og Árni Bergmann. Blátt og rautt. Bernska og unglingsár í tveim heimum. Mál og menning. Reykjavík, 1986.

53 Jónína Michaelsdóttir. Þuríður Pálsdóttir. Líf mitt og gleði. Minningar Þuríðar Pálsdóttur söngkonu. Forlagið. Reykjavík, 1986.

54 Elísabet Þorgeirsdóttir. Í sannleika sagt. Lífssaga Bjarnfríðar Leósdóttur. Forlagið. Reykjavík, 1986.

55 Ingólfur Margeirsson. Ragnar í Smára. Listasafn ASÍ og Lögberg. Reykjavík, 1982.

56 Sveinn Einarsson. Níu ár í neðra. Mynd af Iðnó. Almenna bókafélagið. Reykjavík, 1984.

57 Svavar Gestsson. Magnús Kjartansson. Þeir settu svip á öldina. Íslenskir stjórnmálamenn, bls. 273–287. Sigurður A. Magnússon ritstýrði. Iðunn. Reykjavík, 1983.

58 Bergsteinn Jónsson. Thor Jensen kaupmaður, útgerðarmaður og bóndi. Þeir settu svip á öldina. Íslenskir athafnamenn II, bls. 283–304. Ritstjóri Gils Guðmundsson. Iðunn. Reykjavík, 1988.

59 Sigurður Á. Friðþjófsson. Íslenskir utangarðsunglingar. Vitnisburður úr samtímanum. Forlagið. Reykjavík, 1988.

60 Guðmundur Daníelsson. Á miðjum vegi í mannsaldur. Ólafs saga Ketilssonar. Tákn. Reykjavík, 1988.

## 4. Fræðslutextar

61 Stefán Már Stefánsson. Hlutafélög. Réttarreglur. Hið íslenska bókmenntafélag. Reykjavík, 1985.

62 Bragi Guðmundsson og Gunnar Karlsson. Uppruni nútímans. Íslandssaga frá öndverðri 19. öld til síðari hluta 20. aldar. Bráðabirgðaútgáfa. Mál og menning. Reykjavík, 1986.

63 Jóhannes Nordal. Lífsnauðsyn að brjótast út úr vítahring verðbólgunnar. Fjármálatíðindi. XXX. árg. 2, maí-júlí, bls. 75–87. 1983.

64 Drífa Pálsdóttir. Réttarstaða barna að íslenzkum lögum. Úlfljótur. 4. tbl. 1980, XXXIII. árg., bls. 155–164. Orator, félag laganema, Háskóla Íslands. Reykjavík.

65 Hörður Bergmann (ritstj.). Fréttabréf um vinnuvernd, 1. tbl. 1. árg. Vinnueftirlit ríkisins. Reykjavík, 1984.

66 Þór Whitehead. Stríð fyrir ströndum. Ísland í síðari heimsstyrjöld. Almenna bókafélagið. Reykjavík, 1985.

67 Jón Hnefill Aðalsteinsson. Þjóðtrú og þjóðfræði. Iðunn. Reykjavík, 1985.

68 Sigfús Jónsson. Sjávarútvegur Íslendinga á tuttugustu öld. Hið íslenska bókmenntafélag. Reykjavík, 1984.

69 Aldís Guðmundsdóttir og Jörgen Pind. Sálfræði. Hugur og hátterni. Mál og menning. Reykjavík, 1981.

70 Steinar J. Lúðvíksson. Árbók Íslands. Hvað gerðist á Íslandi 1982. Örn og Örlygur. Reykjavík, 1983.

71 Sigríður Theodórsdóttir og Sigurgeir Jónsson. Efnafræði fyrir menntaskóla. 1. hefti. Menntaskólinn við Hamrahlíð. 1982.

72 Skýrsla um starfsemi Hafrannsóknastofnunarinnar 1983. Hafrannsóknir, 29. hefti, bls. 20–40. Ritstj.: Guðni Þorsteinsson, Eiríkur Þ. Einarsson. Hafrannsóknastofnun. Reykjavík, 1984.

73 Ingimar Jónsson. Heilsufræði. Iðunn. Reykjavík, 1980.

74 Þór Jakobsson. Um heima og geima. Prentsmiðjan Leiftur. Reykjavík, 1983. [Efni frá 1980 og 1982.]

75 Hreggviður Norðdahl og Þorleifur Einarsson. Hörfun jökla og sjávarstöðubreytingar í ísaldarlok á Austfjörðum. Náttúrufræðingurinn, 58. árg. 2. tbl., bls. 59–80. Reykjavík, 1988.

76 Jóhann Sigurjónsson. Stöðvun hvalveiða og áætlun um eflingu hvalrannsókna árin 1986–1989. Sjávarfréttir, 4. tbl. 13. árg., bls. 19–26. 1985. Hvalveiðar við Ísland og fyrirhuguð stöðvun þeirra árið 1986. Jólablað Víkings 1984.

77 Bragi Árnason. Rannsóknir á íslenskum orkulindum. Þættir úr rannsóknarsögu háskólakennara. Í hlutarins eðli. Afmælisrit til heiðurs Þorbirni Sigurgeirssyni prófessor, bls. 167–190. Ritstjóri Þorsteinn I. Sigfússon. Menningarsjóður. 1987.

78 Einar H. Guðmundsson. Sprengistjarnan SN1987A. Jakob Yngvason og Þorsteinn Vilhjálmsson ritstj.: Eðlisfræði á Íslandi IV. Ráðstefna Eðlisfræðifélags Íslands í Munaðarnesi 1.–2. október 1988. Sérhefti af 8. árgangi Fréttabréfs Eðlisfræðifélags Íslands, bls. 11–30. Eðlisfræðifélag Íslands. 1989.

79 Jón Jónsson. Jarðsaga svæðisins milli Selvogsgötu og Þrengsla. Ferðafélag Íslands, Árbók 1985, bls. 63–82.

80 Halldór Kristjánsson. Tölvur og hugbúnaður. Rit fyrir almenning og skóla. Rit Tölvu- og verkfræðiþjónustunnar 1. Reykjavík, 1987.

## 5. Barna- og unglingabækur

81 Rúnar Ármann Arthúrsson. Er andi í glasinu? Svart á hvítu. Reykjavík, 1987.

82 Andrés Indriðason. Með stjörnur í augum. Mál og menning. Reykjavík, 1986.

83 Guðmundur Ólafsson. Emil og Skundi. Vaka. Reykjavík, 1986.

84 Hrafnhildur Valgarðsdóttir. Leðurjakkar og spariskór. Æskan. 1987.

85 Guðlaug Richter. Jóra og ég. Mál og menning. Reykjavík, 1988.

86 Ármann Kr. Einarsson. Hundakofi í paradís. Námsgagnastofnun. 1985.

87 Páll H. Jónsson. Lambadrengur. Iðunn. Reykjavík, 1981.

88 Þórður Helgason. Ég er kölluð Lilla. Námsgagnastofnun. Reykjavík, 1988. [2. útg. (1. útg. 1985)]

89 Kristín Steinsdóttir. Fallin spýta. Vaka-Helgafell. Reykjavík, 1988.

90 Sigurbjörn Einarsson. Af hverju, afi? Talað við börn í jólahug. Skálholt. Reykjavík, 1983.

91 Gripe, Maria. Sesselja Agnes –- undarleg saga. Vilborg Dagbjartsdóttir þýddi. Mál og menning. 1985.

92 Carroll, Lewis. Alís í Undralandi. Ingunn E. Thorarensen íslenskaði. Fjölvaútgáfan. Reykjavík, 1983.

93 Hjorth, Vigdis. Birkir + Anna sönn ást. Ingibjörg Hafstað og Þuríður Jóhannsdóttir þýddu. Mál og menning. Reykjavík, 1986.

94 Carpelan, Bo. Boginn. Sagan af sumri sem var engu líkt. Gunnar Stefánsson þýddi. Iðunn. Reykjavík, 1982.

95 Nöstlinger, Christine. Jói og unglingaveikin. Jórunn Sigurðardóttir þýddi. Mál og menning. Reykjavík, 1988.

96 Dixon, Franklin W. Hardy-bræður Frank og Jói. Lykill galdramannsins. Drengjasaga. Eiríkur Baldvinsson þýddi. Prentsmiðjan Leiftur. Reykjavík, 1983.

97 Winberg, Anna-Greta. Ég er kölluð Ninna. Völundur Jónsson þýddi. Iðunn. Reykjavík, 1980.

98 Lindgren, Astrid. Ronja ræningjadóttir. Þorleifur Hauksson þýddi. Mál og menning. 1985. [2. útg. (1. útg. 1981)]

99 Lewis, C. S. Sigling Dagfara. Kristín R. Thorlacius íslenskaði. Almenna Bókafélagið. Reykjavík, 1986.

100 Röhrig, Tilman. Ekkert stríð. Þorvaldur Kristinsson þýddi. Forlagið. Reykjavík, 1985.

### 3.1.3 Breytingar á textum

Textarnir voru lesnir yfir af nákvæmni og ótvíræðar ritvillur leiðréttar. Hins vegar var óvenjulegum rithætti, sem rekja má til sérvisku höfunda, ekki breytt að öðru leyti en því sem greint er frá hér á eftir.

Fyrir greiningu voru gerðar eftirfarandi breytingar á textunum:

1. Greinarmerki voru fjarlægð nema þau sem eru orðbundin. Með orðbundnum greinarmerkjum er átt við greinarmerki sem eru hluti af orði, eins og t.d. bandstrik í *Austur-Grænlands*, *BASIC-forrit*.
2. Hástaf var breytt í samsvarandi lágstaf í upphafsorði setningar nema í þeim tilvikum þar sem venja er að rita ætíð hástaf í upphafi orðs, svo sem gildir t.d. um sérnöfn. Einnig var hástöfum breytt í lágstafi í orðum rituðum með hástöfum til áherslu, svo sem í fyrirsögnum.
3. Ritun orða sem innihéldu bókstafinn ‘z’ var breytt til samræmis við gildandi stafsetningarreglur, þ.e.a.s. ‘z’ var breytt í ‘s’ nema þar sem ‘z’ er leyfileg samkvæmt reglunum. Í fjórum tilvikum féll ‘t’ niður um leið: *breytzt* → *breyst* (tvívegis), *bætzt* → *bæst*, *setzt* → *sest*.

### 3.1.4 Aðgreining lesmálsorða

Hugtakið *lesmálsorð* nær til samfelldrar raðar af bókstöfum og/eða tölustöfum og táknum sem aðgreind eru með stafbili eða greinarmerkjum. Oftast nær veldur engum erfiðleikum að greina eitt lesmálsorð frá öðru, en þó koma upp álitamál þegar tölur og tákn ýmiss konar eiga í hlut. Í fyrsta kafla þessarar bókar (bls. 3–554) sést með hvaða hætti lesmálsorðum hefur verið haldið aðgreindum. Hér má nefna eftirfarandi atriði:

1. „Venjuleg“ orð teljast augljóslega sérstök lesmálsorð, eins og t.d. *og*, *maður*, *telja* o.s.frv.

2. Töluorð teljast einnig lesmálsorð og er ætíð reynt að fella eins langan stafastreng undir töluorðið og hægt er. Þannig fá plúsar, mínusar og prósentumerki, svo dæmi séu tekin, að fylgja lesmálsorðinu. *10* og *10.* og *10%* eru þá ólík lesmálsorð.
3. Samkvæmt þeirri reglu að láta töluorð taka til eins langs stafastrengs og hægt er, er $0{,}7 \times 10^6$ talið eitt lesmálsorð; sama gildir um *10–15, 1900–1930* o.s.frv.
4. Blendingar tölustafa, bókstafa og annarra rittákna sem eiga saman sem ein heild teljast einnig sérstök lesmálsorð: $Al_2O_3H_3$, $H_2O$, $(d/2c)(mc^2E)^2$, $km^2$, $Mg(OH)^2$.
5. Það sem orkar e.t.v. helst tvímælis í flokkuninni er að stærðfræðijöfnur voru einnig flokkaðar sem stök lesmálsorð: $\delta = \div 94‰$. Slíkar jöfnur eru hins vegar fátíðar í könnuninni.
6. Skammstafanir eru greindar sem eitt lesmálsorð eða fleiri eftir því hvernig lesið er úr þeim að jafnaði. Þó eru skammstafanir mælieininga ávallt greindar sem aðeins eitt lesmálsorð. Nokkur dæmi:

   *BSRB* er lesið *„bé-ess-err-bé“* og greint sem eitt lesmálsorð *BSRB*.

   *H.Í.* er lesið *„Háskóli(-a) Íslands“* og greint sem tvö lesmálsorð: *H.* og *Í.* Þau eru síðan felld undir flettiorðin *háskóli* og *Ísland*.

   Skammstöfunin *e.t.v.* er lesin *„ef til vill“* og greind sem þrjú lesmálsorð: *e.*, *t.* og *v.* Þau eru síðan felld undir flettiorðin *ef*, *til* og *vilja*.

   Skammstöfunin *etv.* er lesin *„ef til vill“* og greind sem þrjú lesmálsorð: *e*, *t* og *v*.

   Mælieiningin *km/s* er lesin *„kílómetrar á sekúndu“* en samt greind sem aðeins eitt lesmálsorð: *km/s*.
7. Ritháttur orða í textunum réð einnig nokkru um skil milli einstakra lesmálsorða.

### 3.1.5 Greining orða

Áður en sjálf málfræðigreiningin hófst var textinn klofinn í lesmálsorð og var einu lesmálsorði komið fyrir í hverri línu. Greiningin fól í sér að færður var greiningarstrengur ásamt flettimynd við hvert lesmálsorð.

Með *greiningarstreng* lesmálsorðs er átt við runu af bókstöfum og e.t.v. einum tölustaf sem auðkenna orðflokk orðsins, beygingarmynd þess og stundum fleiri greiningaratriði (sbr. grein 5.1). Sem dæmi má taka greiningu á upphafi eins textans:

*Ég gleymi örugglega aldrei árinu 1980. Þá gerðist allt.*

Niðurstaða greiningarinnar er sem hér segir:

```
f p 1 e n     ég            ég
s g f n e 1   gleymi        gleyma
a - a         örugglega     örugglega
a - a         aldrei        aldrei
n h e þ g     árinu         ár
t             1980          1980

a - a         þá            þá
s m f þ e 3   gerðist       gera
f o h e n     allt          allur
```

Fremst er greiningarstrengurinn, síðan kemur lesmálsorðið og að lokum flettimyndin. Auð lína er milli setninga.

#### 3.1.5.1 Vélræn greining

Í upphafi könnunarinnar var fyrirsjáanlegt að mestur tími og fyrirhöfn myndi fara í greiningu orðanna. Því var ákveðið að flýta fyrir þessum þætti verksins með því að beita vélrænni greiningu með hjálp tölvutækninnar að svo miklu leyti sem vænlegt þætti. Hagkvæmni þess að beita vélrænni greiningu, þótt hún gefi ekki nærri því alltaf rétta niðurstöðu, byggist á því að léttara er að leiðrétta hana handvirkt en framkvæma greininguna að öllu leyti handvirkt.

Vélræna greiningin á fyrstu fimmtíu textunum byggðist á greiningu í eldri orðtíðnikönnun (Friðrik Magnússon 1988) sem náði til rúmlega 54.000 lesmálsorða. Við vélræna greiningu síðustu fimmtíu textanna var hins vegar hægt að byggja á leiðréttri greiningu fyrstu fimmtíu textanna með rúmlega 250.000 lesmálsorðum.

Auk þess að nýta niðurstöðu fyrri greiningar við vélrænu greininguna er tekið mið af því að af orðmynd lesmálsorðs má oft ráða beygingarmynd þess þótt sjaldan sé það ótvírætt. Í sumum tilvikum getur síðari hluti orðs ákvarðað orðflokk þess og beygingarmynd jafnvel þótt fyrri hluti orðsins sé óþekktur. Nýta má vitneskju um endingar við vélræna greiningu á orðum sem ekki hafa komið fyrir áður.

Við vélrænu greininguna er einnig beitt fjöldamörgum einföldum reglum um samband orða í sömu setningu, svo sem reglum um fallstjórn og sambeygingu. Slíkar reglur hjálpa einnig til við greiningu á áður óþekktum orðum.

Endanleg niðurstaða vélrænu greiningarinnar byggist á líkindareikningi, en greiningarforritið reiknar út hvaða reglur eru uppfylltar og hverjar ekki og niðurstaðan ræðst síðan af því.

Það er ýmsum vandkvæðum bundið að leggja mat á það hversu nákvæm vélræna greiningin er, m.a. vegna þess að einstök greiningaratriði voru stundum endurskoðuð í ljósi fenginnar reynslu. Þó er óhætt að fullyrða

að yfir 80% lesmálsorðanna greinast að öllu leyti rétt við vélrænu greininguna, bæði hvað varðar greiningarstreng og flettimynd. (Nánari greinargerð um hina vélrænu greiningu er að finna hjá Stefáni Briem 1990.)

#### 3.1.5.2 Greining yfirfarin og leiðrétt

Eftir vélrænu greininguna á hverjum texta var niðurstaða hennar yfirfarin og leiðrétt og bætt við greiningu á orðum sem vélræna greiningin hafði skilið við ógreind. Þar er einnig greitt úr fjölmörgum vafaatriðum sem upp koma við greininguna og er nánari grein gerð fyrir því í kaflanum um málfræðigreininguna sem fer hér á eftir.

# 4 Málfræðigreiningin

Til þess að unnt sé að kanna tíðni málfræðiatriða og flettiorða er ítarleg málfræðigreining nauðsynleg. Í þessum kafla verður fjallað um þá málfræðigreiningu sem liggur að baki flestum tíðnitölum í þessari bók. Þau málfræðiatriði sem tilgreind eru við lesmálsorðin verða talin upp og fjallað verður um þau vandamál og vafaatriði sem upp hafa komið við greininguna. Fyrst verður fjallað um orðflokkagreininguna í grein 4.1, síðan um málfræðiatriði einstakra orðflokka í grein 4.2 og loks um flettimyndir í grein 4.3.

## 4.1 Orðflokkagreining

Gerður er greinarmunur á átta orðflokkum í þessari könnun: nafnorðum, lýsingarorðum, fornöfnum, lausum greini, töluorðum, sögnum, atviksorðum og samtengingum. Þau orð sem ekki geta talist til þessara orðflokka falla síðan í tvo flokka, annars vegar erlend orð og hins vegar ógreind orð. Helstu frávik frá hefðbundinni íslenskri orðflokkagreiningu (þ.e. þeirri sem kennd hefur verið í skólum hér á landi) eru þau að forsetningar, upphrópanir og nafnháttarmerki eru ekki talin sjálfstæðir orðflokkar. Að þessu verður vikið nánar hér á eftir. Ýmis vafaatriði hafa komið upp við orðflokkagreininguna og eru flest þeirra vel þekkt úr kennslubókum í orðflokkagreiningu og því óþarfi að rekja þau öll nákvæmlega hér. Þó verður ekki undan því vikist að geta helstu vafaatriða og nefna þær meginreglur sem hafðar hafa verið til hliðsjónar við orðflokkagreininguna.

Þótt hér verði nefndar ýmsar málfræðilegar ástæður fyrir greiningunni má ekki gleyma því að ýmsar leiðir hafa verið valdar af hagnýtum ástæðum til að auðvelda greininguna. Ef tvær leiðir eru mögulegar í greiningu (t.d. orðflokkagreiningu) vegur það þungt ef önnur er einfaldari og krefst minni vinnu.

### 4.1.1 Forsetningar og atviksorð

Í þessari könnun eru forsetningar ekki taldar sjálfstæður orðflokkur heldur teljast þær til atviksorða. Í hefðbundinni orðflokkagreiningu er gerður greinarmunur á forsetningum sem stýra falli og atviksorðum sem stýra

yfirleitt ekki falli. Þessir orðflokkar skarast hins vegar mjög mikið eins og fram kemur hjá Birni Guðfinnssyni (1939:99): „Nú geta flestar forsetningar orðið að atviksorðum, og fjölmörg atviksorð verða að forsetningum. Fer þetta eftir *stöðu* og *notkun* orðanna. Gilda ákveðnar reglur um það, hvenær smáorð er *forsetning* og hvenær *atviksorð*. **Forsetning telst það því aðeins, að það stýri falli**.“ Á hinn bóginn þarf smáorð ekki að vera forsetning þótt það stýri falli: „En nú geta sum *atviksorð* einnig *stýrt föllum*, án þess að þau verði að *forsetningum*“ (s.st.). Aftur á móti verða forsetningar „að *atviksorðum*, þegar *fallorð þeirra falla brott*“ (s.st.). Forsetningar verða sem sagt alltaf að atviksorðum þegar þær stýra ekki falli en atviksorð verða aðeins stundum að forsetningum þegar þau stýra falli.

Af þeim 47 orðum sem Björn Guðfinnsson (1939:99) telur til forsetninga koma 46 fyrir í þessari orðtíðnikönnun. Af þessum 46 orðum geta 37 einnig verið atviksorð þegar þau stýra ekki falli, en aðeins níu stýra alltaf falli. Þau eru *án, gagnvart, gegnt, gegnum, handa, mót, sakir, sökum* og *umfram*. Öll eru þessi orð tiltölulega sjaldgæf og saman eru þau ekki nema tæplega 1% af lesmálsorðafjölda forsetninganna 46. Allar algengustu forsetningarnar geta líka verið atviksorð.

Vegna þessarar miklu skörunar forsetninga og atviksorða þótti rétt að endurskoða orðflokkagreiningu forsetninga, sérstaklega með hliðsjón af öðrum orðflokkum. Í kennslubók Björns Guðfinnssonar (1939:125–6) eru nefndir sex orðflokkar, að frátöldum forsetningum, sem geta stýrt falli: áhrifssagnir, nafnorð, fornöfn, lýsingarorð, töluorð og atviksorð. Engir þessara orðflokka (nema sum atviksorð) verða að öðrum orðflokkum þegar þeir stýra falli. Með öðrum orðum, fallstjórn ræður því aldrei hvaða orðflokki orð tilheyrir nema þegar um er að ræða forsetningar og atviksorð. Sagnir halda áfram að vera sagnir hvort sem þær stýra falli eða ekki. *Setja* er sama orðið og tilheyrir sama orðflokki hvort sem það stýrir falli eða ekki. Á sama hátt verður því haldið fram hér að *til* sé sama orðið og tilheyri sama orðflokki hvort sem það stýrir falli eða ekki. Því er hér ekki gerður greinarmunur á atviksorðum og forsetningum í orðflokkagreiningunni heldur verða þau orð sem venjan er að telja til forsetninga talin með atviksorðum. Með því móti fæst einnig betri yfirsýn yfir það hversu oft þessi orð stýra falli og hversu oft þau gera það ekki.

### 4.1.2 Upphrópanir

Hér eru upphrópanir ekki taldar sérstakur orðflokkur eins og í hefðbundinni orðflokkagreiningu heldur eru þær taldar með atviksorðum, enda geta sum atviksorð stundum staðið sem upphrópanir. Til dæmis er *æ* upphrópun í setningunni *Æ, þessi fjandans höfuðverkur*, en venjulegt atviksorð í setningunni *Vegna minnkandi eiginfjármögnunar hafa þeir orðið æ háðari lánsfjármögnun*.

### 4.1.3 Nafnháttarmerki

Nafnháttarmerkið svokallaða er hér greint sem samtenging enda er erfitt að réttlæta þann greinarmun sem gerður hefur verið á því og skýringartengingunni *að* í hefðbundinni orðflokkagreiningu.

### 4.1.4 Tilvísunartengingar

Tilvísunarorðin *sem* og *er*, sem talin eru til fornafna í hefðbundinni orðflokkagreiningu, eru hér talin til samtenginga í samræmi við stöðu þeirra og hegðun. Um rök fyrir þessari greiningu, sjá Höskuld Þráinsson (1980).

### 4.1.5 Lýsingarorð og atviksorð

Það eru ekki aðeins forsetningar sem skarast við atviksorð. Lýsingarorð í hvorugkyni eintölu nefnifalli (eða þolfalli sem alltaf er eins og nefnifallið) eru talin verða að atviksorðum við vissar aðstæður og er yfirleitt skorið úr um orðflokkinn á grundvelli stöðu orðanna og hlutverks. Lýsingarorð standa í ákveðnu falli og eiga við fallorð en atviksorð standa yfirleitt með sögnum, lýsingarorðum eða öðrum atviksorðum og hafa ekkert fall. Þannig má nefna sem dæmi að hvorugkynsmyndin *fast* er greind sem atviksorð í *Markúsína hélt fast við hugmynd sína*, en sem lýsingarorð í *hann hafði komið því í fast horf*. Hér er slíkri greiningu fylgt enda er hægt að stigbreyta sum þessara orða þegar þau eru í stöðu og hlutverki atviksorða og fá þau þá miðstigsendingu atviksorða en ekki lýsingarorða: *hann grípur fastar um pokann*. Fleiri dæmi um atviksorð af þessu tagi eru: *hann reykti pípu allmikið, eldsnöggt greip hún það upp, langt á undan, barst því fljótt fiskifréttin, dálítið fleiri, mikið veik*.

Á hinn bóginn eru lýsingarorð í þágufalli eintölu hvorugkyns ekki greind sem atviksorð jafnvel þótt þau gegni hlutverki atviksorðs, t.d. *löngu síðar, miklu skemmtilegra, nógu vel, skömmu áður*.

### 4.1.6 Lýsingarháttur þátíðar og lýsingarorð

Oft er erfitt að gera greinarmun á lýsingarhætti þátíðar af sögn og lýsingarorði. Þeirri reglu er þó fylgt hér að greina slíkar orðmyndir sem lýsingarhátt þátíðar af sögn nema augljóslega sé um lýsingarorð að ræða, þ.e. þegar orðið hefur bæði stöðu og hlutverk lýsingarorðs. Sem dæmi má nefna að *þveginn* er greint sem lýsingarorð í *Ég var í hvítum frakka og þveginni skyrtu*, en sem lýsingarháttur þátíðar af sögninni *þvo* í *Æskilegt er að líkaminn sé þveginn vel og rækilega að kvöldi til*.

### 4.1.7 Lýsingarháttur nútíðar

Oft er vafamál hvað gera skuli við svokallaðan lýsingarhátt nútíðar. Hann er greindur sem sagnorð í hefðbundinni orðflokkagreiningu (ef ekki sem nafnorð) en hefur hins vegar sjaldnast stöðu eða hlutverk sagnar heldur mun fremur lýsingarorðs eða jafnvel atviksorðs. Þeirri reglu er fylgt hér að greina lýsingarhátt nútíðar sem lýsingarorð þegar hann hefur stöðu

og hlutverk lýsingarorðs: *fljúgandi hálka, hækkandi kaupgjald, kæfandi skítapest, augun urðu stór og spyrjandi, hann veltist í grasinu hlæjandi*; sem atviksorð þegar hann hefur stöðu og hlutverk atviksorðs: *alveg standandi hissa, þetta er vonandi tímabundið ástand, sagði konan afsakandi, Bennet svaraði því neitandi*; en annars sem sagnorð: *verða miklu ráðandi, fór listræn tjáningarþörf hans sívaxandi.*

### 4.1.8 Raðtölur

Raðtölur eru hér taldar til lýsingarorða en ekki töluorða eins og í hefðbundinni orðflokkagreiningu. Ástæðan er fyrst og fremst sú að ekki hefur reynst mögulegt að greina á milli lýsingarorðsins *fyrri, fyrsti* og raðtölunnar *fyrstur* eins og gert er í hefðbundinni orðflokkagreiningu. Í *Íslenskri orðabók* Menningarsjóðs (1983) er lýsingarorðið *fyrri* sagt merkja „sá af tveim sem á undan fer í röð í tíma eða rúmi“ (bls. 256). Þessi merking á hins vegar varla við um efsta stigið *fyrstur*, t.d. í þeim dæmum sem nefnd eru: „*í fyrsta lagi* fyrst; ekki fyrr en; *með fyrsta = í fyrstu* í upphafi“(s.st.). Því er farin sú leið hér að greina efsta stigið *fyrstur, fyrsti* sem sérstakt orð aðskilið frá miðstiginu *fyrri*, og greina allar raðtölur sem lýsingarorð. Einfaldar þetta mjög greiningu á þessum orðum og sparar vinnu við yfirlestur vélrænnar greiningar.

### 4.1.9 Fornöfn og atviksorð

Hér að framan var tekið fram að lýsingarorðsmyndir í hvorugkyni eru greindar sem atviksorð þegar þær gegna hlutverki atviksorðs og standa með sögnum, lýsingarorðum eða atviksorðum, enda er stundum hægt að stigbreyta þessi orð eins og venjuleg atviksorð. Þannig er ekki farið með fornöfn (gjarna óákveðin fornöfn) sem einnig geta gegnt hlutverki atviksorða og staðið með sögnum, lýsingarorðum og atviksorðum, enda er ekki hægt að stigbreyta þau líkt og atviksorðin. Þessar orðmyndir eru því greindar sem fornöfn. Hér koma nokkur dæmi um þessa notkun fornafnanna: *allnokkuð kvenhollur, allt frá landnámsöld, eitthvað fúll, Vaskur skammaðist sín augsýnilega ekkert, honum var nokkuð brugðið, nokkru síðar, öllu heldur, allra best, alls ekki, einna dýpst.* Þó finnast dæmi um að orðmyndir fornafna séu greindar sem atviksorð, t.d. *alls* í merkingunni ‘samtals’, *annars* í merkingunni ‘ella, reyndar’ og *því* í merkingunni ‘þess vegna’, t.d. *alls 240 kr., annars hefði hún ekki verið þarna, ég bið ykkur því að vera viðbúnir.*

### 4.1.10 Einn

Orðið *einn* hefur margbrotna og oft ansi óljósa merkingu. Hér er gerður greinarmunur á þessu orði sem töluorði, óákveðnu fornafni og lýsingarorði. *Einn* er oftast greint sem töluorð, t.d. í *eitt lítið olíumálverk, einn af máttarstólpum þessa litla samfélags, einn eða fleiri, einn og einn bátur*, en sem óákveðið fornafn í merkingunni ‘nokkur, einhver’, t.d. *eitt helsta viðfangsefni hans, svo fátt eitt sé nefnt, einu sinni þekkti ég strák, einna*

*dýpst, eins konar, dag einn*, en annars sem lýsingarorð í merkingunni ‘samur, eins, aleinn, einsamall’ o.fl., svo sem: *allt í einu, það eina sem hún var í, hégómi einn, ýmist einn eða með öðrum, þarna býr hann einn í nýlegu einbýlishúsi*. Það verður þó að viðurkenna að oft er erfitt að greina á milli þessara hlutverka orðsins.

### 4.1.11 Atviksorð, lýsingarorð og nafnorð

Hér verða sýnd dæmi um nokkur atviksorð, lýsingarorð og nafnorð sem hafa reynst erfið í orðflokkagreiningunni. Þessi orð eru af ýmsu tagi og oft eru ekki önnur rök fyrir greiningunni en tilfinning manna fyrir orðflokkunum. Aðalreglan er sú að greina orðin í orðflokka eftir hlutverki þeirra og setningarlegri stöðu en þó eru orðmyndir greindar sem nafnorð ef hægt er með góðu móti að finna þeim einhverja flettimynd í nefnifalli. Þó má nefna að *konar* er greint sem nafnorð þótt ekki hafi tekist að finna því neina aðra og betri flettimynd en *konar*.

1. Eftirtalin orð eru greind sem atviksorð (þegar þau eru í stöðu atviksorðs og gegna hlutverki þess): *allskostar, annarsstaðar, álengdar, ámóta, einhvernveginn, einhversstaðar, einusinni, enganveginn, hástöfum, hinsvegar, loksins, mestanpart, miðsvæðis, neðanjarðar, neinsstaðar, nokkurntíma, samdægurs, snemmsumars, stundum, sumsstaðar, umhverfis, utanbókar, vitaskuld, víðsvegar*.
2. Þessi orð eru greind sem lýsingarorð (þegar þau eru í stöðu lýsingarorðs og gegna hlutverki þess): *allskonar, ámóta, bévítans, einhverskonar, elskuhjartans, fernskonar, gamaldags, hverskonar, stóreflis, sundurorða, tvímastra, utangátta, þesskonar, örmagna*.
3. Dæmi um orðmyndir sem greindar eru sem hvorugkynsnafnorð (jafnvel þótt þau séu í stöðu atviksorðs eða lýsingarorðs): *eftirlætis, fádæma, kvenkyns, ósköp, ótal, síðdegis, undra*; sem karlkynsnafnorð: *afburða, andskotans, aumingja, djöfuls, fjandans, fjárans, konar, veslings*; og sem kvenkynsnafnorð: *alvöru, elsku, furðu*.

### 4.1.12 Erlend orð

Þau erlend orð sem koma fyrir í samfelldum íslenskum texta eru greind eins og um íslensk orð sé að ræða (sjá þó grein 4.2.1 um greiningu nafnorða hér á eftir). Ef þau koma ekki fyrir í samfelldum íslenskum texta, eða eru innskot, eru orðin greind sem erlend orð. Þetta geta verið:

1. heilar setningar: *Home is where the heart is, Et sygt samfundssystem skaber i første omgang syge oprörere, Omnis mundi creatura quasi liber et pictura nobis est in speculum*,
2. upphrópanir og ávarpsorð: *love, adieu, salut, cara, au revoir*,
3. langir titlar og heiti: *Analytische Sozialpsychologie und Gesellschaftsteorie, Un hiver à Majorque, over farmasöytiske spesialpreparater, De rerum natura, Groundwater systems in Iceland traced by deuterium*,

4. bein ræða á erlendu máli: *„D'accord. Je sais. Mais nous," stamaði Houston*,
5. þeir hlutar erlendra sérnafna sem ritaðir eru með lágum upphafsstaf, svo sem *de, da, dal, aga, di* og *von* í þessum nöfnum: *Pierre de St. Laurent, Jóakim Soares da Cunha, Castel dal Monte, Abdi aga, Accademia di Belle Arti, Joachim von Ribbentrop*.

### 4.1.13 Ógreind orð

Þau orð sem hvorki er hægt að flokka í orðflokka né geta beinlínis talist af erlendum toga eru látin vera ógreind. Hér er aðallega um að ræða staka bókstafi eins og *a, b, c* í upptalningu, skammstafanir sem ekki hefur tekist að lesa úr, efnaformúlur eins og $Al_2O_3H_3$ og sértákn ýmiss konar, svo sem =, ∴, → og *.

## 4.2 Málfræðiatriði einstakra orðflokka

Hér verður greint frá því hvaða málfræðiatriði eru tilgreind með hverju lesmálsorði og hvaða erfiðleikar hafa komið upp við greininguna. Aðalreglan við málfræðigreininguna er sú að reyna að fullgreina hvert einasta lesmálsorð. Eins og við er að búast hefur það ekki tekist algerlega en þau orð sem ekki hefur tekist að greina lenda annaðhvort í flokki erlendra orða eða ógreindra (sjá um þau hér að framan). Við hvern orðflokk er birt tafla sem sýnir þau tákn sem notuð eru við málfræðigreininguna. Í grein 5.1.1 er heildartafla yfir greiningaratriði allra orðflokka.

### 4.2.1 Nafnorð

| 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 |

Við sérhvert nafnorð er tilgreint kyn, tala og fall svo og hvort það hefur viðskeyttan greini og hvort um er að ræða mannsnafn, örnefni eða sérnafn af öðrum toga.

Kynin eru þrjú: karlkyn, kvenkyn og hvorugkyn, en þar sem ekki hefur tekist að greina kyn orðanna, þrátt fyrir að þau komi fyrir í setningarlegu samhengi og hægt sé að greina fall þeirra og tölu, eru þau höfð ókyngreind en greind að öðru leyti eins og venjuleg nafnorð. Þessi nafnorð eru þrenns konar: 1) 'Erlend örnefni og önnur sérnöfn, og jafnvel nokkur erlend mannanöfn þar sem kyn persónanna kemur ekki fram. Dæmi um slík orð eru: *Vesterbrogade, Boston, Politiken, Telegraph, CIA, Karamsín, Friedman*. 2) 'Erlendar slettur eins og: *rugby, aspargus, slapstick, sauna*. 3) 'Frumefnaskammstafanir eins og: *Li, Na, S, Ar*.

Þá má nefna að íslensk ættarnöfn og erlend skírnar- og ættarnöfn eru greind eftir náttúrulegu kyni. Þannig er *Parker* og *Norðmann* greint sem karlkyn í *Parker var samanrekinn maður* og *til Jóns Norðmann píanóleikara*, en sem kvenkyn í *sagði frú Parker* og *Jórunn Norðmann og börn hennar*.

Ennfremur eru skammstafanir á borð við *ASÍ, BHM, MA* kyngreindar samkvæmt aðalorði nafnliðarins. Þannig eru til dæmis *ASÍ, BHM, KFUM* greind sem hvorugkyn eins og aðalorð þessara nafnliða *alþýðusamband, bandalag, félag*, en *MA, MR, MS* eru greind sem karlkyn þar sem *menntaskóli* er aðalorð þessara nafnliða.

Engir teljandi erfiðleikar hafa komið í ljós við greiningu nafnorða í eintölu og fleirtölu og viðskeyttur greinir er ennþá auðveldari viðfangs. Fallgreiningin er auðvitað heldur flóknari en þar hafa einstök dæmi fyrst og fremst valdið erfiðleikum. Slíkt kemur aðallega fyrir í föstum orðasamböndum eða orðatiltækjum þar sem erfitt er að bæta viðskeyttum greini við nafnorðið eða hafa með því lýsingarorð eða annað ákvæðisorð, sem sýnt getur fallið betur en nafnorðið, eins og *í þokkabót*, svo og þar sem nafnorð eru notuð sem áhersluorð, eins og *ósköp mjó, aumingja konunum dauðbrá, alvöru listamaður*. Þá eru nafnorð flokkuð eftir því hvort þau eru samnöfn, mannanöfn, örnefni eða annars konar sérnöfn (þ.e. sérnöfn sem hvorki teljast mannanöfn né örnefni). Hér verða síðasttöldu flokkarnir þrír nefndir einu nafni sérnöfn til aðgreiningar frá þeim fyrstnefnda, samnöfnum. Notuð hefur verið tiltölulega einföld regla til að greina sérnöfn frá samnöfnum því aðeins nafnorð sem skrifuð eru með háum upphafsstaf eru talin til sérnafna. Með þessa reglu að leiðarljósi hefur reynst auðvelt að greina sérnöfn frá samnöfnum, nema helst í þeim örfáu tilvikum þar sem orðin koma aðeins fyrir í upphafi setninga, og eru þar af leiðandi með háum upphafsstaf, án þess að augljóst sé hvort um samnafn eða sérnafn er að ræða.

Af sérnafnareglunni leiðir að aðeins nafnorð koma til greina sem sérnöfn og einnig að séu nöfn eða heiti gerð af mörgum nafnorðum er hið fyrsta þeirra greint sem sérnafn en hin því aðeins að þau séu skrifuð með háum upphafsstaf. Þannig er aðeins fyrsta orð eftirfarandi samsettra nafna greint sem sérnafn: *Gráskinna hin meiri, Þrepin þrettán, Bolsíur frá bernskutíð*, en öll nafnorð sem hefjast á háum upphafsstaf í þessum: *Ferðabók Eggerts og Bjarna, Bæjarútgerð Hafnarfjarðar, Alþýðusamband Íslands*. Sama gildir um erlend nöfn sem gerð eru af mörgum lesmálsorðum nema hvað þar eru öll orð sem rituð eru með háum upphafsstaf talin til sérnafna hvort sem þau teljast til nafnorða í erlenda tungumálinu eða ekki. Þannig eru öll orð sem rituð eru með háum upphafsstaf í eftirtöldum nöfnum greind sem nafnorð og sérnöfn: *Accademia di Belle Arti, British Museum, Gare de L'Ouest, Jan Mayen, Jóakim Soares da Cunha, Los Angeles, New York, The Daily Telegraph, Walter von Knebel*.

Sérnöfn eru síðan flokkuð í mannanöfn, örnefni og önnur sérnöfn, eins og áður hefur komið fram:

1. Mannanöfn eru nöfn manna, dýra og annarra lifandi vera, svo sem: *Aðalbjörg, Egill, Fúsi, Janet, Skalla-Grímur, Vaskur*.
2. Örnefni eru nöfn landa, borga, fjalla, fljóta, gatna, húsa o.s.frv., svo sem: *Akranes, Bókhlöðustígur, Cambridge, Eldborgarhraun, Engey, Faxaflói, Laxá, Lækjartorg, Neptúnus, Norður-Atlantshaf, Portúgal, Skarðsfjöruviti, Skólavörðuholt, USA, Vesterbrogade, Zimbabwe*.
3. Þau sérnöfn sem hvorki geta talist mannanöfn né örnefni eru greind sem önnur sérnöfn. Dæmi um slík orð eru: *Aflatryggingasjóður, Austfirðingur, Blöndahlskaramella, Bretavinna, Buick, Dani, Eimskipafélag, Framsóknarflokkur, Gestapó, Hátíðarforleikur, Jónsmessa, Kötlugos, Lugerbyssa, Norðurlandaráð, Sogamýrarvagn, Sturlungar, Öskubuskusaga*.

### 4.2.2 Lýsingarorð

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Orðflokkur | l-**lýsingarorð** |
| 2 | Stig | f-frumstig, m-miðstig, e-efsta stig |
| 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 |

Við sérhvert lýsingarorð er tilgreint stig þess og beyging, en auk þess kyn, tala og fall. Stig lýsingarorða eru þrjú: frumstig, miðstig og efsta stig. Beyging lýsingarorða er annaðhvort veik (*hvíta kanínan*) eða sterk (*hvít kanína*), ellegar að lýsingarorðið er óbeygjanlegt, þ.e. form þess breytist ekki við beygingu. Dæmi um óbeygjanleg lýsingarorð eru: *agndofa, farlama, forviða, framandi, frávita, gjaldþrota, hissa, hringlaga, lotningarvekjandi, ómálga, sænskumælandi, yfirþyrmandi, yxna*. Engir teljandi erfiðleikar komu í ljós við greiningu þessara beygingaratriða lýsingarorða.

### 4.2.3 Fornöfn

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Orðflokkur | f -fornafn |
| 2 | Flokkur | a-ábendingarfn., b-óákveðið ábendingarfn., e-eignarfn., o-óákveðið fn., p-persónufn., s-spurnarfn., t-tilvísunarfn. |
| 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 |

Fornöfnum er hér skipt í sjö undirflokka: persónufornöfn, eignarfornöfn, ábendingarfornöfn, spurnarfornöfn, óákveðin fornöfn, óákveðin ábendingarfornöfn og tilvísunarfornafn. Þessi flokkun er frábrugðin hefðbundinni flokkun fornafna á þrjá vegu: Í fyrsta lagi er afturbeygða fornafnið

*sig* ekki talið sérstakur undirflokkur heldur talið með persónufornöfnum (afturbeygt persónufornafn) til samræmis við afturbeygða eignarfornafnið *sinn* sem ætíð hefur verið talið til eignarfornafna. Í öðru lagi eru fornöfnin *hvílíkur, samur, sjálfur, slíkur* og *þvílíkur* talin sérstakur undirflokkur, óákveðin ábendingarfornöfn, en þessi orð hafa oftast verið greind sem nokkurs konar undirflokkur ábendingarfornafna (nema helst *hvílíkur*). Í þriðja lagi er tilvísunarfornafnið hér aðeins eitt, *hver*, en tilvísunarorðin *sem* og *er* eru talin til samtenginga eins og áður hefur komið fram.

Að öðru leyti eru fornöfn greind sem hér segir: Greind er persóna, tala og fall fornafna 1. og 2. persónu, en kyn, tala og fall annarra fornafna.

Helstu erfiðleikar í greiningu fornafna eru fólgnir í skiptingu þeirra í undirflokka; þó nokkur dæmi eru um að sama orðmyndin geti tilheyrt mismunandi fornöfnum. Mest ber á þessu þar sem fornafn 3. persónu er í hvorugkyni eintölu og allri fleirtölunni eins og ábendingarfornafnið *sá*, þ.e. orðmyndirnar *það, því, þess, þeir, þá, þeim, þeirra, þær, þau*. Munur þessara fornafna er einkum fólginn í því að persónufornafnið er oftast nær sjálfstætt en ábendingarfornafnið stendur oftast með öðru fallorði, t.d. nafnorði eða lýsingarorði. Þó getur verið snúið að greina á milli þeirra, einkum þegar fornafnið stendur ekki með öðru fallorði. Erfitt er að gefa ákveðnar reglur um greinarmun fornafnanna í slíkum tilvikum en þó má nefna að oftast er það ábendingarfornafnið en ekki persónufornafnið sem tekur með sér tilvísunarsetningu. Til dæmis er *það* ábendingarfornafn í: *Ég nenni ekki að vera að bera á borð það sem enginn vill borða.*

### 4.2.4 Laus greinir

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 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 |

Laus greinir er aðeins eitt orð í íslensku, *hinn*, og er hann greindur í kyni, tölu og falli. Hann er hér talinn sérstakur orðflokkur eins og jafnan í hefðbundinni orðflokkagreiningu en vel hefði komið til greina að flokka hann með fornöfnum, sem undirflokk fornafna, vegna stöðu hans og hlutverks innan nafnliðarins.

### 4.2.5 Töluorð

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Orðflokkur | t-töluorð |
| 2 | Flokkur | f-frumtala |
| 3 | Kyn | k-karlkyn, v-kvenkyn, h-hvorugkyn |
| 4 | Tala | e-eintala, f-fleirtala |
| 5 | Fall | n-nefnifall, o-þolfall, þ-þágufall, e-eignarfall |

Töluorðum er skipt í tvennt í þessari könnun: Frumtölur og aðrar tölur (eins og áður hefur komið fram eru raðtölur taldar til lýsingarorða). Aðrar tölur eru aðallega þrenns konar:

1. Ártöl, númer og fleiri óbeygjanlegar tölur svo sem *árið 1843, klukkan þrjú, númer þrettán, Ránargata 18, Stöð 5*. Þessar tölur beygjast ekki þótt nafnorðin sem þær standa með geti beygst: árið 1843, árinu 1843, ársins 1843. Þarna er töluorðið alltaf eins, endar á *þrjú*, þótt nafnorðið beygist. Þessi töluorð beygjast því hvorki í kynjum né föllum þótt þau séu að formi til eins og nefnifall/þolfall hvorugkyns samsvarandi frumtalna. Aðrar óbeygjanlegar tölur sem nefna má hér eru kaflanúmer, gráðutölur af ýmsu tagi (*60°C, 273°K, 64°30', -69°202*) og ýmsar tölur sem ekki flokkast annars staðar (*1-7-2-7*).
2. Prósentutölur eins og *10%, 1,8%*. Þessar tölur geta að vísu beygst en þar sem ekki er ótvírætt hvernig lesa á úr prósentumerkinu, *prósent* eða *af hundraði*, eru þær ekki greindar frekar.
3. Fjöldatölur sem standa framan við töluorðin *hundrað* og *þúsund*, svo sem: *tíu þúsund hermenn, fjögur hundruð milljónir líra, tvö hundruð hrjáðir einstaklingar*. Í þessum dæmum eru töluorðin *hundrað* og *þúsund* greind eins og hliðstæð töluorð í sama kyni, tölu og falli og nafnorðin sem þau standa með (hermenn, milljónir, einstaklingar). Hins vegar beygjast fjöldatölurnar ekki í samræmi við nafnorðin: tvö hundruð hrjáðir einstaklingar, um tvö hundruð hrjáða einstaklinga, frá tvö hundruð hrjáðum einstaklingum, til tvö hundruð hrjáðra einstaklinga.

Þessi þrenns konar töluorð eru ekki greind frekar (og fá því eingöngu orðflokkamerkinguna T), en frumtölurnar eru greindar í kynjum, tölum og föllum. Skiptingin á milli eintölu og fleirtölu er að því leyti öðruvísi hjá frumtölum en öðrum fallorðum (nema kannski örfáum nafnorðum) að þær beygjast ekki í eintölu og fleirtölu heldur eru *einn* og samsettar frumtölur sem enda á *einn*, svo sem *tuttuguogeinn, 528.431*, í eintölu, en aðrar frumtölur eru í fleirtölu.

### 4.2.6 Sagnir

| **Dálkur** | **Formdeild** | **Greiningartákn-greiningaratriði** |
|---|---|---|
| 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 |

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Orðflokkur | s-sögn (lýsingarháttur þátíðar) |
| 2 | Mynd | g-germynd, m-miðmynd |
| 3 | Háttur | þ-lýsingarh. þátíðar |
| 4 | Kyn | k-karlkyn, v-kvenkyn, h-hvorugkyn |
| 5 | Tala | e-eintala, f-fleirtala |
| 6 | Fall | n-nefnifall, o-þolfall |

Við sérhverja sögn er tilgreind mynd og háttur. Myndir sagna eru tvær, germynd og miðmynd, og þarfnast þær ekki frekari útskýringa (sjá þó tölulið 10) í grein 4.3 um flettimyndir og flettiorð hér á eftir).

Hættirnir eru sjö: nafnháttur, boðháttur, framsöguháttur, viðtengingarháttur, sagnbót, lýsingarháttur þátíðar og lýsingarháttur nútíðar. Hér að framan var minnst á erfiðleika við að gera greinarmun á lýsingarháttunum og lýsingarorðum (og atviksorðum) og er því óþarfi að endurtaka það hér en hins vegar er rétt að skýra muninn á sagnbót og lýsingarhætti þátíðar.

Gerður er greinarmunur á óbeygjanlegri sagnbót og beygjanlegum lýsingarhætti þátíðar. Sagnirnar *hafa* og *geta* taka með sér sagnbót, eins og í setningunum *þau höfðu ekkert rætt um sumarfríið, hún gat ekki annað en brosað*, en sagnir eins og *vera* og *verða* taka með sér lýsingarhátt þátíðar (t.d. í þolmynd), eins og í setningunum *hann var kominn með herraklippingu, svo var hvískrinu haldið áfram.*

Auk hátta og mynda eru sagnir í persónuháttum (framsöguhætti, viðtengingarhætti og boðhætti) greindar í tíð, tölu og persónu, sagnir í lýsingarhætti þátíðar eru greindar í kyni, tölu og falli, en sagnir í öðrum háttum (nafnhætti, sagnbót og lýsingarhætti nútíðar) eru ekki greindar frekar (nema hvað merkt er við þátíð nafnháttar af sögnunum *munu* og *vilja*).

### 4.2.7 Atviksorð

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 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 |

Hér að framan var því lýst hversu mjög atviksorð skarast við aðra orðflokka og nefndir helstu erfiðleikar við afmörkun orðflokksins. Þegar orðflokkagreiningu er lokið eru atviksorðin ekki til mikilla vandræða í greiningunni.

Í fyrsta lagi eru miðstig og efsta stig atviksorða merkt sérstaklega. Í öðru lagi eru atviksorðin greind eftir fallstjórn eða flokki í fimm hópa: 1) þau sem ekki stýra falli, 2) þau sem stýra þolfalli, 3) þau sem stýra þágufalli, 4) þau sem stýra eignarfalli og 5) upphrópanir. Hér koma reyndar upp ýmis vafaatriði þar sem ekki er alltaf ljóst hvort orðin stýra falli eða ekki vegna

þess að oft hafa nafnliðirnir verið fluttir frá fallvaldinum eða einfaldlega felldir brott. Sem dæmi má nefna að í tilvísunarsetningum er algengt að sá nafnliður sem atviksorðið stýrir falli á hafi verið felldur brott: *veitingahúsið sem hún hafði fyrst farið inn á_____*. Þarna, og í fleiri svipuðum tilvikum þar sem augljóst er að nafnliðurinn hefur verið felldur brott, er atviksorðið talið stýra falli (þolfalli í þessu tilviki) og er það reyndar í samræmi við hefðbundna greiningu nema hvað þá er talið að „forsetningin“ stýri falli „tilvísunarfornafnsins“.

### 4.2.8 Samtengingar

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Orðflokkur | c-samtenging |
| 2 | Flokkur | n-nafnháttarmerki, t-tilvísunartenging |

Samtengingar eru ekki greindar á annan hátt en þann að merkt er sérstaklega við tilvísunartengingarnar *sem* og *er* og einnig við nafnháttarmerkið *að*.

### 4.2.9 Erlend orð og ógreind

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 1 | Flokkur | e-erlent orð |
| 1 | Flokkur | x-ógreint orð |

Eins og áður hefur komið fram eru þau orð sem ekki geta talist til orðflokkanna átta ýmist greind sem erlend orð eða ógreind. Þessir tveir flokkar eru ekki greindir frekar.

## 4.3 Flettimyndir og flettiorð

Í þessari könnun er tilgangurinn ekki aðeins sá að kanna tíðni orðmynda og málfræðiatriða heldur einnig tíðni flettiorða. Til þess að það sé mögulegt er nauðsynlegt að halda saman mismunandi orðmyndum sama flettiorðs. Það er gert á þann hátt að færa upp sérstaka flettimynd við hvert lesmálsorð sem sýnir hvaða flettiorði lesmálsorðið tilheyrir. Hér verður heitið flettimynd einskorðað við þessa merkingu: mynd sem sýnir hvaða flettiorði tiltekið lesmálsorð tilheyrir. Flettimynd fallorða er í karlkyni, eintölu, nefnifalli og flettimynd sagna er í nafnhætti. Þannig er *unglingur* flettimynd lesmálsorðsins *unglingarnir*, *fullorðinn* flettimynd lesmálsorðsins *fullorðna*, *ákveða* flettimynd lesmálsorðsins *ákváðu*, o.s.frv.

Hér á eftir verður sagt frá þeim meginreglum sem fylgt hefur verið við uppsetningu flettimynda og helstu vandamálum og vafaatriðum sem komið hafa í ljós, en segja má að þau séu aðallega tvenns konar: a) „mismunandi“ orð geta haft sömu flettimynd og b) „sama“ orðið getur haft mismunandi flettimyndir.

1) Vandamálin við uppsetningu flettimynda eru bundin aðgreiningu orða innan orðflokka. Þegar búið er að ákvarða orðflokk lesmálsorðs skiptir ekki máli þótt flettimynd orðsins sé hin sama og flettimynd orðs úr öðrum orðflokki vegna þess að flettiorð úr ólíkum orðflokkum geta aldrei talist sama flettiorðið. Með öðrum orðum: þótt flettimynd kvenkynsnafnorðsins *saga* sé hin sama og sagnarinnar *saga* teljast þessi orð ekki til sama flettiorðsins vegna þess að orðflokkurinn er ekki hinn sami.

Á sama hátt geta nafnorð af ólíkum kynjum ekki talist til sama flettiorðsins þótt flettimynd þeirra sé hin sama. Þannig eru kvenkynsnafnorðið *ár / árin* og hvorugkynsnafnorðið *ár / árið* mismunandi flettiorð þótt flettimyndin sé hin sama, *ár*. Sama er að segja um *egg / eggin (kvk) –- egg / eggið (hk), leiði / leiðinn (kk) –- leiði / leiðið (hk), reiði / reiðinn (kk) –- reiði / reiðin (kvk)*, og fleiri slík pör.

Að öðru leyti gildir sú regla að orð í sama orðflokki, og nafnorð í sama kyni, sem hafa sömu flettimynd og sömu beygingu eru talin sama flettiorðið jafnvel þótt merking þeirra sé ólík. Merkingin ein dugir því aldrei til að greina á milli flettiorða. Dæmi um þetta er hvorugkynsnafnorðið *lag* sem talið er sama flettiorðið hver svo sem merking þess er, t.d. í eftirfarandi setningahlutum: *allt í lagi, einkum og sér í lagi, í mesta lagi, hvort í sínu lagi, í fyrsta lagi, utan laga og réttar, eins og jaxl í laginu, í efsta lagi vökvans, að slá hann út af laginu, fallegasta lagið sem hún kunni.*

2) Orð sem hafa sömu flettimynd en mismunandi beygingu og merkingu eru aðgreind með merktum flettimyndum. Þessi orð eru sagnirnar *bera / bar –- bera[1] / beraði, brenna / brann –- brenna[1] / brenndi, elda / eldi / elti –- elda[1] / eldaði, enda / endaði –- enda[1] / enti, heyja / háði –- heyja[1] / heyjaði, hverfa / hvarf –- hverfa[1] / hverfði, meina / meinti –- meina[1] / meinaði, muna / mundi –- muna[1] / munaði, mæla / mældi –- mæla[1] / mælti, renna / rann –- renna[1] / renndi, róa / reri –- róa[1] / róaði, skella / skall –- skella[1] / skellti, sleppa / slapp –- sleppa[1] / sleppti, smella / small –- smella[1] / smellti, sökkva / sökk –- sökkva[1] / sökkti, vara / varar / varaði –- vara[1] / varir / varði –- vara[2] / varir / varaði, velta / valt –- velta[1] / velti, æja / áði –- æja[1] / æjaði*, og nafnorðið *arður / arð –- arður[1] / arður.*

Orð sem hafa sömu flettimynd og sömu merkingu en mismunandi beygingu eru á hinn bóginn ekki aðgreind heldur höfð undir sömu flettimynd. Þar má nefna eftirnafn eins og *Gunnarsson* sem er sama flettiorðið hvort sem það er af erlendum toga og eins í öllum föllum eða er íslenskt föðurnafn og fær beygingarendingar í þágufalli og eignarfalli.

3) Nafnorð sem hafa ólíka merkingu í eintölu og fleirtölu eru færð undir sömu flettimyndina í eintölu. Dæmi um slíkt eru: *átak / átök, fat / föt, gagn / gögn, lag / lög.*

4) Orð sem stafsett eru á mismunandi hátt (annaðhvort þar sem ýmist er stafsett eftir framburði eða ekki, eða þar sem tvenns konar stafsetning er hugsanleg þótt hún sýni ekki endilega framburðarmun) eru sameinuð undir

eina og sömu flettimynd og ræður hefðbundin stafsetning oftast vali flettimyndar. Hér koma nokkur dæmi um þetta og er hefðbundna stafsetningin (þ.e. flettimyndin) höfð framan við skástrik en hin óhefðbundna aftan við: *almennilega / alminlega, bleia / bleyja, Dyrhólaey / Dyrhóley, ég / eg, fjörutíu / fjörtíu, franskbrauð / fransbrauð, gifs / gips, grafkyrr / grafkjurr, grenja / gðenja, guð / gvuð, Grímur / Gríímur, indíáni / indjáni, kannski / kannske, koníak / coníak, kröftuglega / kröftulega, lukt / lugt, lúterskur / lútherskur, milljón / miljón, ofboðslega / obboðslega, orrusta / orusta, praktískur / praktiskur, predika / prédika, sautján / seytján, skipta / skifta, stríður / strýður.*

Sömu sögu er að segja af orðum sem koma ýmist fyrir með bandstriki eða án þess; þau eru yfirleitt færð undir flettimynd án striks: *bang / bang-bang-bang, dúdúfugl / dúdú-fugl, Kaspían / Ka-ka-kaspían, mahóníborð / mahoní-borð, nettógjaldeyriseign / nettó-gjaldeyriseign.* Þetta gildir einnig um skástrik þannig að *svart/hvítur* er fært undir flettimynd án skástriks: *svarthvítur.*

Undantekningar frá ofangreindri reglu eru þrenns konar:

1. Karlkynsnafnorðið *kall* og kvenkynsnafnorðið *kelling* eru höfð undir þessum flettimyndum en ekki færð undir flettimyndirnar *karl* og *kerling.* Sama gildir um ýmsar samsetningar með *kall* sem síðari lið: *gormakall, hórkall, járnkall, kerfiskall, Kínakall, sprellikall, tíkall, tóbakskall, tröllkall, öskukall.* Svo vill reyndar til að engin þessara samsetninga kemur fyrir með *-karl* sem síðari lið.
2. Nokkur orð sem víkja mjög verulega frá hefðbundinni stafsetningu eru ekki færð undir „rétta“ flettimynd, enda er vafamál hvort það teljast sömu orðin þegar stafsetningunni hefur verið breytt svo mikið: *oná, oneftir, oní, soldið, soldill.*
3. Stafsetning mannanafna er látin ráða flettimynd þeirra. Eftirtaldar nafnatvenndir (eða -þrenndir) eru því dæmi um ólík flettiorð: *Alfred / Alfreð, Annie / Anný, Bennet / Bennett, Carol / Carole, Carl / Karl, Eggerts / Eggerz, Halfdan / Hálfdan / Hálfdán, Hendriks / Hendrix, Janice / Janis, Joseph / Jósef / Jóseph, Julia / Júlía, Lúðvig / Lúðvík, Soffía / Sofia / Sofía, Solveig / Sólveig, Valdemar / Valdimar.*

Undantekningar frá þessari undantekningu eru tvær: lesmálsorðin *Jesum* og *Jésú* eru færð undir flettimyndina *Jesús*; einnig er lesmálsorðið *Zoega* fært undir flettimyndina *Zoëga.*

5) Samsett orð sem eru eins að öðru leyti en því að fyrri liðurinn er annaðhvort stofn, eignarfall (eintölu eða fleirtölu) eða stofn ásamt bandstaf eru aðskilin sem tvö ólík flettiorð. Dæmi: *aðgerðalykill / aðgerðarlykill, annarsstaðar / annarstaðar, augnalok / augnlok, byggingaiðnaður / byggingariðnaður, fiskimjöl / fiskmjöl, kristalglas / kristalsglas, náttúrlega / náttúrulega, rannsóknarstofa / rannsóknastofa, siðferðilegur / siðferðislegur, tilviljanakenndur / tilviljunarkenndur.*

6) Til eru orð sömu merkingar sem hafa mismunandi flettimyndir og ólíka beygingu að undanskildum einhverjum sameiginlegum beygingarmyndum þar sem orðin verða ekki aðgreind. Dæmi um slík orð eru *hólmi* og *hólmur* sem hafa að nokkru leyti sömu merkingu en beygjast aðeins að hluta til eins og hafa ekki sömu flettimynd. Hins vegar er fleirtala þessara orða alveg eins, þannig að orðmynd eins og *hólmum* er ekki með góðu móti hægt að telja til annars orðsins fremur en hins vegna svipaðrar merkingar þeirra, allra síst ef hún er eina dæmið sem finnst í textanum. Orð af þessu tagi eru hér færð undir sameiginlega flettimynd með svigum og/eða skástriki þar sem reynt er að sýna báðar flettimyndirnar í einni. Dæmi um slíkar flettimyndir eru: *dollar(i), ey(ja), éta/eta, flott(ur), hólmi/-ur, kærleiki/-ur, lær(i), meiðsl(i), meir(a), mey(ja)/mær, reip(i), smíð(i), smyrsl(i), systkin(i), umlukinn/-luktur*.

Þetta er því aðeins gert að fyrir komi sameiginleg beygingarmynd. Ef svo er ekki (t.d. ef eingöngu kemur fyrir ein beygingarmynd sem aðeins getur tilheyrt öðru flettiorðinu) er aðeins færð upp einföld flettimynd. Hér koma nokkur dæmi um slíkt og eru orðmyndirnar sem fyrir koma sýndar innan sviga: *dvergasmíð (dvergasmíð), hljóðfærasmíði (hljóðfærasmíði), kvartdollar (kvartdollar), Akurey (Akureyjar), haldreipi (haldreipi), Gunnarshólmi (Gunnarshólma)*. Einnig eru dæmi um að tvær eða fleiri orðmyndir komi fyrir og engin þeirra geti átt við tvær flettimyndir. Þá eru færðar upp tvær mismunandi flettimyndir og flettiorðin því tvö. Dæmi um slíkt eru (orðmyndirnar sem fyrir koma eru innan sviga): *fáleiki (fáleikinn) –- fáleikur (fáleik), heilagleiki (heilagleiki) –- heilagleikur (heilagleik), myndugleiki (myndugleika) –- myndugleikur (myndugleik), sannleiki (sannleiki, sannleika, sannleikann, sannleikanum, sannleikans) –- sannleikur (sannleikur, sannleikurinn, sannleik), sjúkleiki (sjúkleika) –- sjúkleikur (sjúkleikur), veruleiki (veruleiki, veruleikinn, veruleika, veruleikann, veruleikanum, veruleikans) –- veruleikur (veruleik)*. Rétt er að benda á að ofangreindar orðmyndir eru allar í eintölu.

7) Mannanöfn og örnefni eru færð undir sérstaka flettimynd þótt samsvarandi samnafn finnist í textunum, eða sé til. Sem dæmi má nefna að mannsnafnið *Álfur* er haft undir flettimynd með háum upphafsstaf, aðskilið frá samnafninu *álfur*. Dæmi um mannanöfn af þessu tagi eru: *Bára, Björn, Fjóla, Gestur, Lína, Þröstur*. Dæmi um örnefni eru: *Bakki, Brekka, Nes, Tangi, Vík, Þrengsli*.

Ef sama nafnorðið getur bæði verið mannsnafn og örnefni er það fært undir sömu flettimynd, eins og t.d. *Venus, Virginía*, og ef samsvarandi samnafn kemur fyrir er það undir sérstakri flettimynd, aðskilið frá mannsnafninu og örnefninu. Dæmi um slíkt eru: *Jökull –- jökull, Ormur –- ormur, Steinn –- steinn, Úlfur –- úlfur*. Orð af þessu tagi eru þó því aðeins talin til sama flettiorðs ef um sama kyn er að ræða, samanber það sem sagt var í tölulið 1) um að nafnorð af ólíkum kynjum geti ekki talist til sama flettiorðsins þótt flettimynd þeirra sé hin sama. Þannig eru eftirtalin nafnorð talin mis-

munandi flettiorð: Berg (mannsnafn í karlkyni) –- Berg (örnefni í hvorugkyni) –- berg (samnafn í hvorugkyni), Jean (mannsnafn í karlkyni) –- Jean (ókyngreint örnefni).

8) Þau sérnöfn sem hvorki eru mannanöfn né örnefni eru hér flokkuð undir heitinu „önnur sérnöfn“ (ýmist nefnd svo eða einfaldlega „sérnöfn“ hér á eftir). Þau eru því aðeins færð undir sérstaka flettimynd að ekki finnist samsvarandi samnafn, mannsnafn eða örnefni í textunum, annars eru þau færð undir sömu flettimynd og samnafnið, mannsnafnið eða örnefnið. Dæmi um sérnöfn af þessu tagi sem færð eru undir sérstaka flettimynd (með háum upphafsstaf) eru: *Aflatryggingasjóður, Björgunarfélag, Hafnfirðingur, Íslandsklukka*. Dæmi um sérnöfn af þessu tagi sem færð eru undir flettimynd samsvarandi samnafns eru: *Andi, Fæðing, Lúðrasveit, Sjúkrahús, Tilraun, Ævisaga*. Dæmi um sérnöfn af þessu tagi sem færð eru undir flettimynd samsvarandi mannsnafns eru: *Freyja, Galileo, Iðunn, Snorri, Trausti*. Dæmi um sérnöfn af þessu tagi sem færð eru undir flettimynd samsvarandi örnefnis eru: *Hekla, Helgafell, Keflavík, Reykjanes, Vatnajökull*.

Ef dæmi finnast um að samsvarandi orð geti verið sérnafn, samnafn og mannsnafn og/eða örnefni er sérnafnið fært undir flettimynd samnafnsins. Þannig eru sérnöfnin *Drottning* og *Frón* færð undir flettimyndirnar *drottning* og *frón* þótt dæmi finnist um samsvarandi örnefni, og sérnöfnin *Dísa, Haukur* og *Hlíf* færð undir flettimyndirnar *dísa, haukur* og *hlíf* þótt dæmi um samsvarandi mannanöfn komi fyrir.

9) Í grein 4.2.1 kom fram að kyn íslenskra ættarnafna og erlendra skírnar- og ættarnafna sé greint eftir náttúrulegu kyni. Einnig kom fram í tölulið 1) hér að framan að nafnorð af ólíkum kynjum geti ekki talist til sama flettiorðsins þótt flettimynd þeirra sé hin sama. Þetta veldur því að mannsnafn sem bæði kemur fyrir sem karlmannsnafn og kvenmannsnafn telst tvö flettiorð. Dæmi um slík nöfn eru: *Alice, Brown, Cardone, Eggerz, George, Hansen, Kranz, Liang, Möller, Nielsen, Norðmann, Parker, Thorarensen, Thorsteinsson, Willard*.

10) Ekki er færð upp sérstök flettimynd sagna í miðmynd nema samsvarandi germynd sé ekki til. Sem dæmi má nefna að færð er upp flettimyndin *koma* við orðmyndina *komast* og flettimyndin *minna* við orðmyndina *minnast*, en hins vegar er flettimyndin *heppnast* færð upp við orðmyndina *heppnast* og flettimyndin *farnast* við orðmyndina *farnaðist* þar sem sagnirnar **heppna* og **farna* eru ekki til.

11) Töluorð hafa nokkra sérstöðu meðal orðflokkanna vegna þess að hægt er að tákna tölur á a.m.k. þrennan hátt: 1) með tölustöfum (*21*), 2) með rómverskum tölum (*XXI*) og 3) með orðum sem sýna hvernig lesið er úr tölunum (*tuttugu og einn*). Þetta veldur því að þegar texti er bútaður niður í stök lesmálsorð greinast samsett töluorð í mörg lesmálsorð þegar þau eru táknuð með orðum (*tuttugu og fimm* greinist t.d. í þrjú lesmálsorð) en ekki þegar þau eru táknuð með tölustöfum (t.d. *23*) eða rómverskum

tölum (t.d. *XXIX*). Af þessu leiðir að tíðni töluorða veltur að nokkru leyti á því hvernig samsettu töluorðin eru táknuð.

Hér eru bókstafatölur, tölustafatölur og rómverskar tölur ekki settar undir sömu flettimyndina og því eru *6, VI* og *sex* þrjú mismunandi flettiorð. Þá eru ógreind töluorð (sem aðeins fá orðflokkamerkinguna T) færð undir sömu flettimynd og samsvarandi frumtala ef hún kemur fyrir.

12) Svokallaðar fleiryrtar samtengingar (*til þess að, vegna þess að, svo að* o.fl.) eru hér bútaðar niður í einstök lesmálsorð og þau síðan greind hvert út af fyrir sig (sem atviksorð, fornöfn, samtengingar o.s.frv.) og færð hvert undir sína flettimynd. Stundum eru fleiryrtar samtengingar þó ritaðar í einu orði og eru þær þá greindar sem eitt lesmálsorð (og sú orðmynd færð upp sem flettimynd) og orðflokkagreindar sem samtengingar, t.d. *einsog, þarsem, tilað, þvíað, áðuren*.

13) Stundum er illmögulegt að finna flettimynd í karlkyni, eintölu, nefnifalli þegar um lýsingarorð er að ræða, og er þá ekki átt við óbeygjanleg lýsingarorð (sjá grein 4.2.2 um málfræðiatriði lýsingarorða hér að framan). Þetta eru lýsingarorð sem hafa „frosið“ í einstökum orðatiltækjum eða orðasamböndum, og/eða eru notuð þar sem ekkert samræmi fæst við annað fallorð í kyni, tölu og falli, eða þá að samræmið er alltaf við fallsetningu (sem alltaf er í hvorugkyni eintölu) eða *það*. Þessi lýsingarorð koma aðeins fyrir í hvorugkyni, eintölu, nefnifalli eða þolfalli og er sú orðmynd þá færð upp sem flettimynd. Dæmi um þetta eru orðmyndir eins og *annt, ábótavant, flökurt, óhætt, ómótt* og *viðvart*, í setningum eins og: *hann lét sér mjög annt um fjölskyldu sína, bændur eru hvattir til að lagfæra strax það sem ábótavant kann að reynast, mér er alltaf flökurt, það er óhætt að treysta Fríðu gömlu, þá verður mér ómótt og ég svitna allur, þá verður að gera ábótanum viðvart*.

Helsti munur á þessum lýsingarorðum og óbeygjanlegum lýsingarorðum er sá að þau síðarnefndu geta yfirleitt staðið hliðstæð með nafnorðum í ýmsum kynjum, tölum og föllum (en líta þó alltaf eins út), en þau fyrrnefndu geta ekki staðið hliðstæð með nafnorðum (**ábótavant drasl*, **flökurt barn*, **ólíft hús*), en eru oftast sagnfyllingar með aukafallsfrumlögum og laga sig því ekki að þeim í kyni, tölu og falli, eða þá að nafnliðirnir sem þau laga sig að eru alltaf í hvorugkyni, eintölu, nefnifalli.

## 5 Um einstaka kafla bókarinnar

Í þessum hluta er að finna yfirlit um kafla í meginmáli bókarinnar (bls. 3 og áfram). Fjallað verður um efnivið hvers kafla og greint frá því með hvaða hætti hann var unninn úr niðurstöðum tíðnikönnunarinnar. Kaflarnir eru tölusettir frá 1–14 og er vitnað til kaflanúmers með tölustaf innan hornklofa.

## 5.1 Flettiorð og greiningarmyndir [1]

### 5.1.1 Heildarskrá um orðaforða könnunarinnar

Í fyrsta kafla er heildarskrá um orðaforða tíðnikönnunarinnar. Hún er samtals 552 bls. að lengd. Í kaflanum er að finna öll flettiorð sem greind voru í textasýnunum 100 ásamt öllum greiningarmyndum hvers flettiorðs sem dæmi voru um í textunum. Á næstu síðu er stutt sýnishorn úr kaflanum:

| | | |
|---|---|---|
| **fóstra** *no* | **1** | **1** |
| fóstra NVEN | | 1 |
| **fóstra** *so* | **1** | **1** |
| fóstrar SGFNE3 | | 1 |
| **fóstri** *no* | **3** | **4** |
| fóstri NKEN | | 1 |
| fóstra NKEO | | 2 |
| fóstra NKEE | | 1 |
| **fóstur** *no* | **6** | **9** |
| fóstur NHEO | | 5 |
| fóstri NHEÞ | | 1 |
| fósturs NHEE | | 3 |

Í þessu sýni eru fjögur flettiorð, kvenkynsnafnorðið *fóstra*, sögnin *fóstra*, karlkynsnafnorðið *fóstri* og hvorugkynsnafnorðið *fóstur*.

Orðaskráin er þannig úr garði gerð að flettiorðum er raðað í stafrófsröð en síðan er greiningarmyndunum raðað í sérstakri „málfræðiröð“. Þetta sést í fyrrgreindu sýni; af nafnorðinu *fóstri* eru þrjár greiningarmyndir, í nefnifalli, þolfalli og eignarfalli eintölu án greinis. Þeim er raðað í þessa röð eins og sést á síðasta bókstaf greiningarstrengsins þar sem röðin er N (nefnifall), O (þolfall) og E (eignarfall). Gerð er nánari grein fyrir stafrófsröðun þeirri sem fylgt er í bókinni svo og hinni sérstöku röðun málfræðiatriða hér á eftir.

Við hvert flettiorð er tilgreind skammstöfun á orðflokki þess. Þar á eftir fara tvær tölur. Fyrri talan greinir frá því í hve mörgum textum (af 100) flettiorðið kemur fyrir en seinni talan greinir frá heildarfjölda dæma um flettiorðið.

Eftirfarandi skammstafanir eru notaðar um orðflokk flettiorða:

| | |
|---|---|
| *ao* | atviksorð |
| *erl* | erlent orð |
| *fn* | fornafn |
| *gr* | greinir |
| *lo* | lýsingarorð |
| *no* | nafnorð |

| | |
|---|---|
| *ógr* | ógreint orð |
| *so* | sagnorð |
| *st* | samtenging |
| *to* | töluorð |

*Tafla I. Orðflokkaskammstafanir flettiorða.*

Á stöku stað er einnig getið um kyn nafnorða þegar orðið er til í fleiri kynjum en einu. Eru þá notaðar skammstafanirnar *kk* fyrir karlkyn, *kvk* fyrir kvenkyn og *hk* fyrir hvorugkyn. Auk þess eru notaðar skammstafanirnar *ófn*, *sfn* og *tfn* til að greina á milli óákveðins fornafns, spurnarfornafns og tilvísunarfornafns þegar ástæða er til.

Eins og þegar hefur verið vikið að eru notuð formleg einkenni til að greina á milli flettiorða. Mestu máli skiptir auðvitað orðflokkurinn, orð teljast til mismunandi flettiorða ef orðflokkurinn er ekki hinn sami. En auk þess greinir kyn nafnorða milli flettiorða, svo og eðli fornafns. Einnig hefur verið greint á milli sagna sem hafa sömu flettimynd en ólíka beygingu að öðru leyti (sbr. grein 4.3).

Orðmyndir eru felldar undir viðkomandi flettiorð og fylgir hverri orðmynd greiningarstrengur sem sýnir beygingu orðsins. Orðmyndin ásamt greiningarstreng nefnist greiningarmynd eins og fyrr segir. Hverri greiningarmynd fylgir tala og greinir hún frá fjölda dæma sem eru um viðkomandi greiningarmynd í textasafni könnunarinnar.

Greiningarstrengur sá sem fylgir hverri orðmynd er þannig úr garði gerður að málfræðiatriði eru táknuð með eins stafs skammstöfunum. Fyrsti stafur greiningarstrengsins táknar ætíð orðflokkinn. Nafnorð er táknað með N, lýsingarorð með L o.s.frv. Merking táknanna er háð þeim „dálki“ sem þau standa í, og er í töflu ‘II að finna yfirlit um allar þær skammstafanir sem notaðar eru í greiningarstrengjunum. Eins og fyrr segir táknar fyrsti stafur greiningarstrengsins orðflokkinn. Ef fyrsti stafurinn er N er t.d. um nafnorð að ræða. Kyn nafnorðsins er tilgreint í öðrum dálki, talan í þeim þriðja, fallið í fjórða og í fimmta dálki er haft G ef orðið er með viðskeyttum greini. Loks er sjötti stafurinn hafður til að greina á milli ólíkra afbrigða sérnafna (M = mannanöfn, Ö = örnefni og S = önnur sérnöfn).

Reynt hefur verið að hafa skammstafanir gagnsæjar og er því yfirleitt notaður fyrsti stafur í heiti viðkomandi málfræðiatriðis. Þó getur slíkt leitt til árekstra. Þannig er þágufall táknað með Þ og verður því að nota annan staf fyrir þolfall og varð O fyrir valinu.

Svo hugað sé aftur að því sýnishorni úr orðaskránni sem birt er að framan sést að eitt dæmi er um nafnorðið *fóstra* í könnuninni. Það er orðmyndin *fóstra* sem hefur greiningarstrenginn NVEN. Með hliðsjón af töflu ‘II má lesa úr honum á þann hátt að orðmyndin sé nafnorð (N) í kvenkyni (V) eintölu (E) og standi í nefnifalli (N). Eitt dæmi er um sögnina *fóstra* og er það orðmyndin *fóstrar* sem hefur greiningarstrenginn SGFNE3. Þar er um að ræða sögn (S) í germynd (G), framsöguhætti (F) nútíðar (N) eintölu (E) þriðju persónu (3).

Ekki er alltaf hirt um að færa upp greiningarmyndir við flettiorð. Það er t.d. ekki gert þegar í hlut eiga erlend orð eða ógreind, eða ógreind töluorð. Í þeim tilvikum er flettiorðið tilgreint eitt sér. Í einstaka tilvikum hefur verið skotið inn skýringum innan hornklofa við flettiorð þegar ekki er augljóst hvernig á að lesa úr þeim; á það t.d. við um bandstrik í merkingunni *til* (bls. 3). Loks er þess að geta að sum orðin eru svo löng að þurft hefur að skipta þeim á milli lína. Þegar það er gert er línuskiptingarbandi skotið aftan við fyrri hluta orðsins. Ef bandstrik er þar fyrir er það flutt niður í næstu línu með síðari hluta orðsins og má af því ráða hvort bandstrikið er hluti af orðinu eða ekki. Sjá t.d. orðið *tvöþúsund-og-fimmhundruð-krónur* (bls. 484).

| **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**-efsta stig |
| 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**-ábendingarfn., **B**-óákveðið ábendingarfn., **E**-eignarfn., **O**-óákveðið fn., **P**-persónufn., **S**-spurnarfn., **T**-tilvísunarfn. |
| 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 |
| 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. persónal |
| 1 | Orðflokkur | **S-sögn** (lýsingarháttur þátíðar) |
| 2 | Mynd | **G**-germynd, **M**-miðmynd |

| Dálkur | Formdeild | Greiningartákn-greiningaratriði |
|---|---|---|
| 3 | Háttur | **Þ**-lýsingarh. 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 | Flokkur | **X**-ógreint orð |

*Tafla II. Skýringar skammstafana í greiningarstrengjum.*

### 5.1.2 Röðun

#### 5.1.2.1 Stafrófsröð

Frumskilyrði þess að orðabókarnotandi geti fundið það sem hann er að leita að í orðabók er að orðunum sé raðað með einhverjum hætti. Algengast er að raða þeim í stafrófsröð og er þeirri venju einnig fylgt í kaflanum um flettiorð og beygingarmyndir [**1**]. Í öðrum köflum er orðunum stundum raðað í tíðniröð (t.d. [**3**] og [**7**]) eða í stafrófsröð eftir niðurlagi orðanna ([**6**] og [**8**]).

Við fyrstu sýn mætti ætla að ekki þyrfti að eyða mörgum orðum í að greina frá því að orðum í þessari bók sé raðað í stafrófsröð (en þó hefur stafrófsröðun orðið mörgum fræðimönnum tilefni lærðra greina, sbr. t.d. Gavare 1988). Nokkuð fastar venjur hafa mótast um það með hvaða hætti ber að raða orðum í stafrófsröð og höfum við fylgt þeirri hefð sem nú er sem óðast að festast í sessi hérlendis og felur í sér þá nýjung að raða íslenskum broddstöfum sérstaklega (sbr. Baldur Jónsson 1987).

Í orðtíðnibókinni er að finna ýmis tákn sem ekki eru hluti af hinu venjulega stafrófi og því þykir rétt að birta hér töflu um þá röðun sem notuð var. Í meginatriðum er röðunin sú að táknum og tölustöfum er raðað fremst, því næst kemur stafrófið og er broddstöfum haldið sér og raðað næst á eftir skyldum sérhljóða (t.d. fer ‘á’ næst á eftir ‘a’). Ýmsum erlendum stöfum með stafmerkjum (eins og t.d. ‘ä’ og ‘ü’) er raðað milli skylds sérhljóða og broddsérhljóða (‘ä’ lendir þannig á milli ‘a’ og ‘á’).

Við röðun eru hástafir og lágstafir lagðir að jöfnu en ef enginn munur er á orðum annar en sá að annað hefst á hástaf en hitt á samsvarandi lágstaf er hástafsorðinu raðað næst á eftir lágstafsorðinu. Orðaskránum er öllum raðað „að orðabókarhætti“ og er þá einvörðungu tekið tillit til bókstafa og tölustafa við röðun. Önnur tákn (eins og t.d. bandstrik) hafa ekki áhrif á röðun. Því lendir orð eins og *A-bekkur* á milli orðanna *Abdi* og *Aberdeen*

en að öðrum kosti hefði það lent á milli orðanna *A-4* og *a-d* og þessi orð öll hefðu komið næst á eftir *A*.

Öll röðun var framkvæmd í tölvum með `sort`-forriti Unix-stýrikerfisins. Notaðir voru rofarnir `-f` og `-d`. Örfáum táknum þurfti síðan að handraða að lokinni meginröðun þar sem þau eiga sér ekki samsvörun í tölvustafrófinu. Gilti það t.d. um ð og →.

Á næstu blaðsíðu fer tafla sem sýnir röðun allra táknanna sem fyrir koma í bókinni. Lesið er úr töflunni með því að lesa niður eftir hverjum dálki áður en byrjað er á næsta dálki.

| | | | | |
|---|---|---|---|---|
| % | . | 8 | é | r |
| ‰ | : | 9 | f | s |
| & | = | a | g | t |
| ’ | → | à | h | u |
| ( | / | å | i | ü |
| ) | ° | ä | í | ú |
| * | ½ | á | j | v |
| + | 0 | b | k | w |
| ± | 1 | c | l | x |
| , | 2 | d | m | y |
| - | 3 | ð | n | ý |
| – [til] | 4 | ð | o | z |
| – [mínus] | 5 | e | ó | þ |
| ÷ | 6 | è | p | æ |
| × | 7 | ë | q | ö |

*Tafla III. Röðun rittákna í Íslenskri orðtíðnibók.*

#### 5.1.2.2 Röðun greiningarstrengja

Í kaflanum um flettiorð og beygingarmyndir er greiningarmyndum sérstaklega raðað innan hvers flettiorðs og er að mestu fylgt „málfræðilegri röðun“ sem byggist að miklu leyti á hefðbundinni beygingarfræði (sbr. Björn Guðfinnsson 1937).

Málfræðiatriðunum er raðað með þeim hætti sem sýndur er í töflu ‘II. Fyrst er raðað eftir fyrsta bókstaf strengsins, síðan öðrum o.s.frv. Þó verður að gæta að því að þegar bókstöfunum er raðað innbyrðis er ekki fylgt stafrófsröðun þeirra. Svo dæmi sé tekið af nafnorðum þá er flettiorðunum haldið aðgreindum eftir kyni eins og fyrr segir. Síðan er röðunin sú að eintala fer á undan fleirtölu (og er það í samræmi við röð táknanna sem notuð eru til að merkja töluna, E og F). Þegar kemur að falli nafnorðsins er röðin hins vegar sú að fyrst fer nefnifall, síðan þolfall, þá þágufall og loks eignarfall. Hér er röðin ekki í samræmi við röðun þeirra stafa sem tákna föllin því ef henni væri fylgt kæmi eignarfallið

fyrst. Með því að hafa hliðsjón af töflu ‘II er hins vegar hægur vandi að sjá hver röðin er.

Sérstök ástæða er til þess að vekja athygli á því hvernig greiningarstrengur sagna er tilgreindur. Ef sögn er í lýsingarhætti þátíðar er röð greiningaratriða þessi: orðflokkur, mynd, háttur, kyn, tala og fall. Að öðrum kosti er röðin þessi: orðflokkur, mynd, háttur, tíð, tala og persóna. Mynd og háttur koma sem sagt ætíð beint á eftir orðflokksskammstöfun sagna en eftir það skilur leiðir eftir því hvort um lýsingarhátt þátíðar eða annan hátt er að ræða.

## 5.2 Algengustu flettiorð í stafrófsröð [2] og í tíðniröð [3]

### 5.2.1 Efni

Í köflum [**2**] og [**3**] er greint frá niðurstöðum um tíðni flettiorða. Í hinum fyrri eru algengustu flettiorðin í stafrófsröð en í hinum síðari er þeim raðað eftir lækkandi væntingartíðni.

Við hvert flettiorð í [**2**] er getið um röð þess, væntingartíðni, tíðni, dreifingarstuðul, í hvaða textaflokkum orðið er að finna og í hve mörgum textum.

Í [**3**] er aðeins getið um hluta þeirra atriða sem eru tiltekin í [**2**], nánar tiltekið um röð orðsins, orðið sjálft og væntingartíðni.

Rétt er að gera nokkra grein fyrir þeim tölfræðilegu hugtökum sem hér er beitt.

### 5.2.2 Dreifing og röð orða –- tölfræðileg hugtök

Við útreikninga á **dreifingarstuðli** og **væntingartíðni** er fylgt þeirri aðferð sem Sture Allén notar í riti sínu *Nusvensk frekvensordbok 1* (1970). Sú aðferð er sótt til könnunar Alphonse Juillands og E. Chang-Rodriguez, *Frequency dictionary of Spanish words* sem út kom 1964.

Orð og einstakar myndir þeirra eru misjafnlega dreifð eftir textum og textaflokkum. Til þess að lýsa tölulega þessari dreifingu og tilgreina hvaða tíðni er líklegust í einhverjum nýjum texta sem fellur að vali á textum í könnuninni eru notaðar stærðirnar dreifingarstuðull *D* og væntingartíðni *Fv*. Þessar stærðir geta átt við orðmynd, flettiorð, beygingarmynd eða annað slíkt.

Við útreikning á dreifingarstuðli og væntingartíðni eru skilgreindar eftirfarandi stærðir:

| | | |
|---|---|---|
| $n$ | = | fjöldi textahluta |
| $m$ | = | meðaltíðni |
| $x$ | = | tíðni í einstökum textahluta |
| $s$ | = | staðalfrávik |
| $F$ | = | heildartíðni |
| $D$ | = | dreifingarstuðull |
| $Fv$ | = | væntingartíðni |

Útreikningarnir eru síðan framkvæmdir eins og nú skal greint frá. Ef textahlutarnir eru misstórir þarf að umreikna gildin á $x$ með tilliti til stærðar textahlutanna, þ.e. fjölda orða í þeim.

Meðaltíðnin $m$ er reiknuð samkvæmt jöfnunni:

$$m = \frac{\sum x}{n}$$

Staðalfrávik $s$ er reiknað samkvæmt jöfnunni:

$$s = \sqrt{\frac{\sum (x - m)^2}{n}}$$

Dreifingarstuðull $D$ er reiknaður samkvæmt jöfnunni:

$$D = 1 - \frac{s}{m\sqrt{n - 1}}$$

Væntingartíðni $Fv$ er reiknuð samkvæmt jöfnunni:

$$F_v = D \cdot F$$

og gildir þá væntingartíðnin fyrir jafnstóran heildartexta.

Niðurstöður útreikninganna má túlka eins og hér segir. Ef $D = 1$ þá leiðir af því að $Fv = F$ sem sýnir fullkomlega jafna dreifingu. Ef $D = 0$ þá verður $Fv = 0$. Þetta gerist ef orðið er aðeins að finna í einum textahluta. Að öðrum kosti er $D$ alltaf á bilinu $0 < D < 1$.

Með því að reikna út væntingartíðni orða er leitast við að draga upp réttari mynd af orðtíðni en ef einvörðungu væri miðað við heildartíðni. Tíðni orða í tilteknum texta segir í raun ekki til um væntanlega tíðni orða í einhverjum öðrum texta. Með væntingartíðninni er hins vegar hægt að áætla tíðni orða óháð þeim textum sem teknir eru til athugunar hverju sinni.

Í öllum tilvikum er væntingartíðnin lægri en heildartíðnin en það skýrist af því að í öðrum heildartexta koma til sögunnar fjöldamörg ný orð sem fá hlutdeild í heildarfjölda lesmálsorða.

Ef dreifingarstuðullinn er nálægt gildinu 1 og væntingartíðnin því næstum jöfn heildartíðninni merkir það að orðið er álíka algengt í öllum textahlutunum.

Ef dreifingarstuðullinn hefur hins vegar miklu lægra gildi en 1 þá er væntingartíðnin miklu lægri en heildartíðnin og orðið er mjög misdreift eftir textahlutum. Komi orðið aðeins fyrir í einum textahluta er dreifingarstuðullinn 0 og sömuleiðis væntingartíðnin.

Í þessari könnun voru dreifingarstuðull og væntingartíðni reiknuð fyrir öll flettiorðin miðað við textaflokkana fimm. Munurinn á fjölda lesmálsorða í textaflokkunum fimm er svo lítill að hann hefur óveruleg áhrif á gildi dreifingarstuðuls. Því voru gildin á *x* ekki umreiknuð með tilliti til stærðar textaflokka eins og strangt til tekið ætti að gera samkvæmt framansögðu.

**Röð** flettiorðanna er reiknuð út frá væntingartíðninni. Það orð sem fær hæsta væntingartíðni er númer 1 í röðinni, orðið með næsthæstu væntingartíðnina er númer 2 o.s.frv. Fyrir kemur að orð fá sömu væntingartíðni og lenda því á sama stað í röðinni. Svo dæmi sé tekið hafa flettiorðin *losna* og *tólf* sömu væntingartíðni, 37,42. Orðið næst á undan þeim er númer 728 í röðinni þannig að þessi tvö orð skipa sæti 729 og 730. Strangt til tekið er raðnúmer þeirra því 729,5 sem er meðaltal af 729 og 730. Í bókinni eru öll raðnúmer tilgreind í heilum tölum og er þá lækkað niður í lægri töluna en þær að öðru leyti auðkenndar með stjörnu. Í [**2**] og [**3**] fá orðin *losna* og *tólf* því raðnúmerið 729*.

Í kafla [**3**] er flettiorðunum raðað eftir lækkandi væntingartíðni. Hafa ber í huga að væntingartíðnin er prentuð með misjafnlega mörgum aukastöfum og var það fyrst og fremst gert af prenttæknilegum ástæðum. Fjöldi tölustafa í gildunum á væntingartíðni gefur því ekki til kynna nákvæmni gildanna. Fræðileg óvissa á þessum gildum er reyndar mun meiri en sem nemur fjölda tölustafanna.

## 5.3 Flettiorð í einstökum orðflokkum [4] og textaflokkum [5]

Í kafla [**4**] er að finna algengustu flettiorð einstakra orðflokka. Í þessum kafla eru alls átta skrár, ein um hvern orðflokk. Í hverri skrá eru 100 algengustu flettiorð hvers orðflokks, en auk þess eru tekin með 25 algengustu flettiorð hvers textaflokks, jafnvel þótt þau séu ekki meðal 100 algengustu flettiorðanna í könnuninni. Röð orðflokkanna er þessi: nafnorð (598–601), lýsingarorð (602–605), fornöfn og laus greinir (606–607), töluorð (608–611), sagnir (612–615), atviksorð (616–619) og samtengingar(620–621).

Ýmislegt fróðlegt kemur í ljós séu þessar skrár skoðaðar. Svo litið sé á nafnorðin fyrst kemur á daginn að orðið *maður* er algengasta nafnorðið ef teknar eru saman niðurstöður um alla textana 100. Staða þess er þó mismunandi eftir textaflokkum. Í 1. og 2. flokki er orðið í fyrsta sæti, í öðru sæti í 3. flokki og í þriðja sæti í 4. og 5. flokki. Algengasta nafnorðið í 3. og 4. flokki er hins vegar *ár* en *mamma* í 5. flokki. Lýsingarorðið *mikill* er algengasta lýsingarorðið í könnuninni í heild og er einnig algengast í öllum textaflokkum nema í 5. flokki þar sem það er í þriðja sæti. Það er kannski ekki tilviljun að í 5. flokki, barna- og unglingabókum, er *lítill* algengasta lýsingarorðið! Meðal annarra orðflokka er heldur meira samræmi að því er varðar algengustu orð. Orðið *í* er alls staðar í fyrsta sæti atviksorða sem og *einn* meðal töluorða og *og* meðal samtenginga.

Í kafla [**5**] er greint frá 100 algengustu flettiorðum í hverjum textaflokki

óháð orðflokkum. Sögnin *vera* og samtengingin *og* skiptast á að vera í fyrsta sæti í textaflokkunum fimm.

## 5.4 Flettiorð í stafrófsröð eftir niðurlagi [6]

Þegar orðum er raðað í orðabókum er það oftast gert með þeim hætti að fyrst er raðað eftir fyrsta staf í orði, síðan öðrum staf og svo áfram að síðasta staf. Er þá auðvelt að finna orðin eftir upphafi þeirra. Frá sjónarmiði málfræðinga getur einnig verið gagnlegt að raða orðum þannig að orð sem eiga sér sameiginlegar endingar lendi saman í orðaskrám. Þannig má ná saman orðum sem enda á tilteknum seinni lið, t.d. *-legur*, *-heiti* o.s.frv. Í kaflanum Flettiorð í stafrófsröð eftir niðurlagi [**6**] er það einmitt gert.

Í kafla [**6**] er auðvelt að leita uppi flettiorð eftir niðurlagi þeirra. Við hvert orð er getið um tíðni þess í heild í öllum textaflokkum samanlagt. Gæta verður þess þegar skráin er skoðuð að til þess að átta sig á stafrófsröðuninni verður að lesa orðin „frá hægri til vinstri“. Þess vegna eru orðin höfð slétt við hægri brún eins og oft tíðkast þegar tölur eru settar í dálka.

| | | |
|---:|---|---:|
| fjarvera | *no* | 5 |
| vistarvera | *no* | 9 |
| nærvera | *no* | 5 |
| séra | *no* | 45 |
| marséra | *so* | 3 |
| grasséra | *so* | 1 |
| spásséra | *so* | 1 |

Þessi framsetning ætti að auðvelda lesendum að finna þær orðendingar sem þeir hafa áhuga á. Eins og sést af dæminu hér að ofan er auk orðsins sjálfs greint frá orðflokki og tíðni þess.

## 5.5 Orðmyndir [7] og [8]

Í næstu tveim köflum er greint frá orðmyndum. Alls reyndust orðmyndir vera 59.343 að tölu í þessari könnun. Í fyrri kaflanum, Algengustu orðmyndir í tíðniröð [**7**], er að finna orðmyndir sem raðað hefur verið í tíðniröð eftir væntingartíðni. Algengasta orðmyndin reyndist vera *og*.

Í síðari kaflanum er að finna upplýsingar um orðmyndir í stafrófsröð eftir niðurlagi [**8**]. Þessari skrá svipar mjög til sambærilegrar skrár um flettiorð [**6**] nema hvað ekki er getið um orðflokk orðmyndanna því steypt er saman orðmyndum er tilheyra ólíkum orðflokkum. Að öðru leyti skýrir þessi skrá sig sjálf.

## 5.6 Málfræðileg margræðni orðmynda [9]

Hér að framan var lítillega minnst á hina vélrænu greiningu sem notuð var við málfræðigreiningu lesmálsorðanna. Fram kom að árangur greiningarinnar varð um 80% sem teljast má nokkuð gott miðað við að um fyrstu gerð forrits er að ræða. En hvers vegna er erfitt að ná fram greiningu sem er 100% rétt? Meginástæðan fyrir því er sú að lesmálsorðin eru málfræðilega margræð, ekki verður alltaf ráðið af orðinu einu saman hvort þar fari nafnorð eða sögn, hvað þá að hægt sé að ákvarða með óyggjandi hætti fall, persónu eða tölu þess (svo dæmi séu tekin). Þessi þáttur var kannaður sérstaklega og í kafla [**9**] eru birtar upplýsingar um þessa „málfræðilegu margræðni“ orðmynda.

Til þess að safna saman upplýsingum um þetta atriði var hver einstök orðmynd athuguð með tilliti til þess hversu margir ólíkir greiningarstrengir gátu fylgt henni. Í ljós kom að orðmyndin *minni* sýndi mesta fjölbreytni, en alls hlaut hún 24 ólíka greiningarstrengi í þessari könnun. Af þeim voru 17 lýsingarorðsstrengir, 5 nafnorðsstrengir, 1 fornafnsstrengur og 1 sagnarstrengur. Af þessu má sjá að ef reynt er að flokka orðið vélrænt í texta hefur forritið úr 24 kostum að velja a.m.k. Ef ná á góðum árangri í vélrænni orðflokkagreiningu verður því að taka tillit til þess samhengis sem orðin standa ‘í.

Í eftirfarandi töflu er greint frá því hversu margir greiningarstrengir fylgdu einstökum orðmyndum. Taflan sýnir að af 59.343 ólíkum orðmyndum í þessari könnun fá 9.441 eða 15,9% fleiri en einn greiningarstreng og eru því málfræðilega margræðar (auk þess sem nokkur hluti þeirra sem aðeins fengu einn greiningarstreng hefðu vafalítið einnig reynst málfræðilega margræðar ef fleiri dæmi hefðu verið um orðmyndirnar í könnuninni). Í kafla [**9**] eru birtar allar orðmyndir sem áttu sér fleiri en fjóra greiningarstrengi.

| **Fjöldi orðmynda** | **Fjöldi greiningarmynda** | **Fjöldi orðmynda** | **Fjöldi greiningarmynda** |
|---:|---:|---:|---:|
| 1 | 24 | 23 | 11 |
| 1 | 22 | 18 | 10 |
| 1 | 21 | 20 | 9 |
| 1 | 20 | 26 | 8 |
| 2 | 19 | 69 | 7 |
| 5 | 17 | 96 | 6 |
| 2 | 16 | 209 | 5 |
| 4 | 15 | 579 | 4 |
| 7 | 14 | 1.772 | 3 |
| 11 | 13 | 6.586 | 2 |
| 8 | 12 | 49.902 | 1 |

*Tafla IV. Yfirlit um margræðni orðmynda.*

## 5.7 Rittákn, tvístöfungar og þrístöfungar [10], [11] og [12]

Í þrem næstu köflum er greint frá tíðni rittákna, tvístöfunga og þrístöfunga.

Við talningu rittákna er ekki gerður greinarmunur á hástöfum og lágstöfum og hástafir því taldir með í tíðni bókstafanna.

Orðabil og setningarleg greinarmerki eru ekki talin með heldur eingöngu þau tákn sem eru hlutar af orðum, eru orðbundin.

Með þessum hætti fást við talningu 75 mismunandi tákn. Tilgreind er þrenns konar tíðni, í lesmálsorðum, orðmyndum og flettiorðum. Niðurstöður eru sýndar hlið við hlið í skránum til að auðvelda samanburð.

Samanburður leiðir til dæmis í ljós að orð sem hafa að geyma bókstafinn ‘þ’ koma tiltölulega oft fyrir, því að hlutfallsleg tíðni þessa bókstafs í lesmálsorðum, 1,57%, er mun hærri en í orðmyndum og flettiorðum, 0,46%.

Í öðrum töflum yfir rittákn er þeim raðað í tíðniröð auk þess sem greint er frá tíðni táknanna fremst í orðum og aftast í orðum. Þar sést t.d., og kemur naumast á óvart, að bókstafurinn ‘ð’ kemur aldrei fyrir fremst í orði og bókstafurinn ‘þ’ aðeins 7 sinnum aftast í orði. Einnig kemur í ljós að algengustu rittákn fremst í flettiorðum eru ‘s’ (14,54%) og ‘h’ (9,33%) og fellur sú niðurstaða vel að fyrirferð þessara stafkafla í orðabókum.

**Tvístöfungar** og **þrístöfungar** eru orðatvennur eða -þrennur og var sérstaklega hugað að tíðni þeirra í lesmálsorðum, orðmyndum og flettiorðum og einnig gerður greinarmunur eftir stöðu þeirra í orðum. Til þess að finna tvístöfunga í tilteknu orði eru ætíð pöruð saman tvö rittákn í senn. Byrjað er við vinstri brún orðsins og haldið til hægri. Svo dæmi sé tekið eru eftirfarandi tvístöfungar í orðinu *tvístöfungar*:

‘tv’, ‘ví’, ‘ís’, ‘st’, ‘tö’, ‘öf’, ‘fu’, ‘un’, ‘ng’, ‘ga’, ‘ar’

og á sama hátt reynast þrístöfungar vera þessir:

‘tví’, ‘vís’, ‘íst’, ‘stö’, ‘töf’, ‘öfu’, ‘fun’, ‘ung’, ‘nga’, ‘gar’

Rétt er að geta þess að einvörðungu var leitað að tvístöfungum og þrístöfungum sem samanstóðu af tveim eða þrem **bókstöfum**. Við talningu tví- og þrístöfunganna voru allir hástafir færðir til lágstafa. Ef reiknað er með 36 stöfum í íslenska stafrófinu er hugsanlegur fjöldi tvístöfunga 36 × 36 eða 1.296 alls. Af þeim fundust 916 í efniviði könnunarinnar eða 70,7%. Hugsanlegur fjöldi þrístöfunga er 36 × 36 × 36 eða 46.656. Af þeim fundust alls 10.141 eða 21,7%. Greint er frá tíðnitölum um tvístöfunga í kafla [**11**] og er þar getið um tíðni allra tvístöfunga. Þar eru reyndar einnig taldir með nokkrir tvístöfungar þar sem einn eða fleiri stafir eru ekki í íslenska stafrófinu, stafir eins og ‘ä’ og ‘ü’. Alls fundust 30 slíkir tví- og þrístöfungar í könnuninni. Tíðnitölum þrístöfunga eru gerð skil í kafla [**12**]. Vegna fjölda þrístöfunga er þó aðeins greint frá þeim ef hlutfallsleg tíðni þeirra var hærri en 0,02%.

Að lokum er ekki úr vegi að benda á fróðlega grein um notkun þrístöfunga við leit að stafsetningarvillum í tölvusettum enskum textum (MacMahon, Cherry og Morris 1978).

### 5.8 Greiningarstrengir [13]

Næsti kafli sýnir yfirlit um alla greiningarstrengi sem fyrir koma í könnuninni. Alls reyndust þeir vera 621 að tölu. Tíðni þeirra er þó æði breytileg. Algengasti greiningarstrengurinn reyndist vera A - A, en alls fengu 50.282 lesmálsorð þann greiningarstreng. Næstur kom C með 42.491 dæmi en 19 greiningarstrengir komu aðeins fyrir einu sinni hver. Greiningarstrengirnir eru bæði tilgreindir í stafrófsröð og í tíðniröð.

### 5.9 Yfirlitstöflur [14]

Í síðasta kafla bókarinnar er að finna fjölda yfirlitstaflna þar sem dregnar eru saman upplýsingar um könnunina í heild og einstaka orðflokka. Efni taflnanna er skýrt í sérstökum skýringartextum með hverri töflu og er því óþarft að hafa fleiri orð um þær hér.

## 6 Að verkalokum

Eins og fyrr er fram komið á orðtíðnikönnun sú sem greint er frá í þessu riti sér nokkuð langa sögu. Fyrstu drög að henni voru lögð með samstarfi Orðabókarinnar og IBM á Íslandi um gerð villuleitarforrits fyrir ritvinnslu sem hófst árið 1984. Í tengslum við þá vinnu var gerð allrækileg orðtíðnikönnun. Upp úr þeirri könnun varð orðasafn villuleitarforritsins til. Örn Kaldalóns, starfsmaður IBM, og Sigurður Jónsson, fyrrverandi starfsmaður Orðabókarinnar, áttu einnig hlut að þessari fyrstu orðtíðnikönnun stofnunarinnar.

Nokkru eftir að vinnu við villuleitarforritið lauk árið 1985 var ákveðið að halda áfram rannsóknum á orðtíðni, einkum með það fyrir augum að safna upplýsingum um ýmis málfræðileg atriði sem ekki hafði áður verið sinnt í íslenskum orðtíðnirannsóknum. Friðrik Magnússon var ráðinn til þess verks og vann við það á árunum 1986 og 1987 og birtust niðurstöður í grein í tímaritinu *Orð og tunga* **1** árið 1988. Í þeirri grein er sagt frá greiningu á efniviði sem er um 1/10 af þeim efniviði sem lagður var til grundvallar þeirri könnun sem birtist í þessari bók.

Í upphafi árs 1989 var stefnan tekin á útgáfu sérstakrar orðtíðnibókar. Stefán Briem var þá ráðinn til verksins auk Friðriks en Jörgeni Pind falin ritstjórn af hálfu stjórnar Orðabókar Háskólans. Hefur verið unnið sleitulaust við þetta verk síðan.

Við verkalok er rétt og skylt að þakka þeim aðilum sem stutt hafa okkur til þessa verks. Stjórn Orðabókar Háskólans hefur sýnt verki þessu mikinn skilning og áhuga frá upphafi. Einkum hefur formaður stjórnarinnar, Jón G. Friðjónsson, beitt sér ötullega fyrir því. Sérstakar þakkir viljum

við einnig færa Svavari Geststssyni, fyrrverandi menntamálaráðherra, sem veitti Orðabókinni 1.800.000 króna styrk til rannsóknarinnar sem hluta af málræktarátaki menntamálaráðuneytisins árin 1989–1990.

Loks færum við samstarfsmönnum okkar á Orðabók Háskólans þakkir fyrir veitta aðstoð, einkum þeim Birni Þór Svavarssyni sem hélt tölvunum gangandi meðan á þessu stóð, Guðrúnu Kvaran sem safnað hefur efni í textasafn Orðabókarinnar og Jóni Hilmari Jónssyni sem hjálpaði við að greiða úr margvíslegum flækjum við málfræðigreininguna. Jón Hilmar Jónsson, Jón G. Friðjónsson og Svavar Sigmundsson lásu formálann og bentu á ýmis atriði sem betur máttu fara.

*Reykjavík, á aðventu 1991*

Jörgen Pind
Friðrik Magnússon
Stefán Briem

# 7 Ritaskrá

## 7.1 Rit sem getið er í formála

Allén, Sture. 1970. *Nusvensk frekvensordbok baserad på tidningstext. 1. Graford.* Data linguistica 1. Almqvist & Wiksell, Stockholm.

Allén, Sture. 1971. *Nusvensk frekvensordbok baserad på tidningstext. 2. Lemman.* Data linguistica 4. Almqvist & Wiksell, Stockholm.

Allén, Sture. 1972. *Tiotusen i topp.* Ordfrekvenser i tidningstext. Data linguistica 6. Almqvist & Wiksell, Stockholm.

Ársæll Sigurðsson. 1938. Um rannsókn á tíðni orða. *Menntamál* 11:96–102.

Ársæll Sigurðsson. 1940. Algengustu orðmyndir málsins og stafsetningarkennslan. *Menntamál* 13:8–42.

Baldur Jónsson. 1975. *Tíðni orða í Hreiðrinu.* Tilraunaverkefni í máltölvun. Unnið í samvinnu við Reiknistofu Raunvísindastofnunar háskólans. 1. Orð í stafrófsröð eftir upphafi þeirra. 2. Orð í stafrófsröð eftir niðurlagi þeirra. 3. Orð í röð eftir lækkandi tíðni. Rannsóknastofnun í norrænum málvísindum, Reykjavík.

Baldur Jónsson. 1978. *Orðstöðulykill að Hreiðrinu.* Háskóli Íslands, Reykjavík.

Baldur Jónsson. 1987. Íslenska stafrófið. Ólafur Halldórsson [ritstj.]. *Móðurmálið: Fjórtán erindi um vanda íslenskrar tungu á vorum dögum.* Vísindafélag Íslendinga, Reykjavík.

Baldur Jónsson. 1990. Orðtalning í eddukvæðum Konungsbókar. *Gripla* VII:169–177.

Baldur Jónsson, Björn Ellertsson og Sven Þ. Sigurðsson. 1980. *Tölvukönnun á tíðni orða og stafa í íslenskum texta.* Raunvísindastofnun Háskólans, Reykjavík.

Baldur Pálsson. 1990. Biblían frá A til Ö: Orðstöðulykill að Biblíunni, útgáfu 1981. Gunnlaugur A. Jónsson [ritstj.].

*Biblíuþýðingar í sögu og samtíð. Studia theologica islandica* 4:31–38.

Björn Guðfinnsson. 1937. *Íslenzk málfræði handa skólum og útvarpi.* Ríkisútvarpið, Reykjavík.

Eiríkur Rögnvaldsson. 1990. Orðstöðulykill Íslendinga sagna. *Skáldskaparmál: Tímarit um íslenskar bókmenntir fyrri alda* I:54–61.

Engwall, Gunnel. 1984. *Vocabulaire du roman franç*ais (1962–1968). Dictionnaire des fréquences. Data linguistica 17. Almqvist & Wiksell, Stockholm.

Francis, F. W. og H. Kuvčera. 1982. *Frequency analysis of English usage.* Houghton Mifflin, Boston.

Friðrik Magnússon. 1988. Hvað er títt? Tíðnikönnun Orðabókar Háskólans. *Orð og tunga* 1:1–49.

Garside, R., G. Leech og G. Sampson [ritstj.]. 1987. *The computational analysis of English: A corpus-based approach.* Longman, London.

Gavare, Rolf. 1988. Alphabetical ordering in a lexicological perspective. *Studies in computer-aided lexicology.* Data Linguistica 18:63–102. Almqvist & Wiksell, Stockholm.

Höskuldur Þráinsson. 1980. Tilvísunarfornöfn? *Íslenskt mál* 2:53–96.
Indriði Gíslason og Sigríður Valgeirsdóttir. 1979. Könnun á tíðni viðtengingarháttar í þátíð. *Íslenskt mál* 1:107–121.
*Íslensk orðabók handa skólum og almenningi.* 1983. Árni Böðvarsson [ritstj.]. Önnur útgáfa aukin og bætt. Bókaútgáfa Menningarsjóðs, Reykjavík.
Johansson, S. og K. Hofland. 1989a. *Frequency analysis of English vocabulary and grammar based on the LOB corpus. Volume 1: Tag frequencies and word frequencies.* Clarendon Press, Oxford.
Johansson, S. og K. Hofland. 1989b. *Frequency analysis of English vocabulary and grammar based on the LOB corpus. Volume 2: Tag combinations and word combinations.* Clarendon Press, Oxford.
Kaeding. F. W. 1897–1898. Häufigkeitswörterbuch der deutschen Sprache. Berlin.
Landau, S. I. 1989. *Dictionaries: The art and craft of lexicography.* Cambridge University Press, Cambridge.
McIlroy, M. D. 1982. Development of a spelling list. *IEEE transactions on communications*, COM-30:91–99.
MacMahon, L. E., L. L. Cherry og R. Morris. 1978. Unix time-sharing system: Statistical text processing. *The Bell System Technical Journal*, July-August, 2137–2154.
Stefán Briem. 1990. Automatisk morfologisk analyse af islandsk tekst. Jörgen Pind og Eiríkur Rögnvaldsson [ritstj.]. *Papers from the Seventh Scandinavian Conference of Computational Linguistics Reykjavík 1989*:3–13. Institute of Lexicography, Institute of Linguistics, Reykjavík.

## 7.2 Önnur rit sem stuðst var við

Aho, A. V., B. W. Kernighan og P. J. Weinberger. 1988. *The AWK programming language.* Addison-Wesley Publishing Company, Reading, Massachusetts.
Ellegård, A. 1978. *The syntactic structure of English texts: a computer-based study of four kinds of text in the Brown University Corpus.* Gothenburg Studies in English, 43. Gautaborg.
Haugen, E. 1942. *Norwegian word studies 1–2.* The University of Wisconsin Press. Madison, Wisconsin.
Kernighan, B. W. og R. Pike. 1984. *The UNIX programming environment.* Prentice-Hall, Englewood Cliffs, New Jersey.
Knuth, D. E. 1984. *The TEXbook.* Addison-Wesley Publishing Company, Reading, Massachusetts.
Lancashire, I. og W. McCarty. 1988. *The humanities computing yearbook 1988.* Clarendon Press, Oxford.
Wall, L. og R. L. Schwartz. 1991. *Programming perl.* O'Reilly \& Associates, Sebastopol, California.
Wichura, M. J. 1988. *The TABLEmanual.* Personal TEX, Mill Valley, California.