Stofnun Árna Magnússonar í íslenskum fræðum
Orðabók Háskólans

Sigrún Helgadóttir
sigruhel@hi.is

Mörkun texta og markaðar málheildir

Máltækniverkefni Orðabókar Háskólans

Máltækni í mótun
Hugvísindaþing 14.3.2009

Mörkun íslensks texta

- Með mörkun (e. *tagging*) er átt við það að merkja orð í samfelldum texta á kerfisbundinn hátt, t.d. með málfræðilegum upplýsingum, nefnimynd (e. *lemma*) orðsins og upplýsingum um setningafræðilegt hlutverk.
- Oft er orðið *mark* notað um málfræðilegt mark. Málfræðilegt mark er greiningarstrengur sem er tengdur orði í texta og segir til um orðflokk orðsins og önnur málfræðileg atriði, t.d. kyn, tölu og fall fallorða og persónu, tölu og tíð sagna.
- Dæmi:

orð **mark** *nefnimynd*	
ég **fp1en** *ég*	**f** fornafn, **p** persónufornafn, **1** fyrsta persóna, **e** eintala, **n** táknar nefnifall
sagði **sfg1eþ** *segja*	**s** sagnorð, **f** framsöguháttur, **g** germynd, **1** fyrsta persóna, **e** eintala og **þ** þátíð.

Aðferðir við (málfræðilega) mörkun

- Elsta aðferð er handvirk greining – mjög tímafrek
- Vélrænar aðferðir
  - **regluaðferðir** (e. *rule based methods*) – orðasafn notað til þess að merkja sérhvert orð í texta með öllum hugsanlegum greiningarstrengjum, reglur byggðar á málfræði hvers tungumáls notaðar til þess að skera úr um hvaða greiningarstrengur er réttur – forrit sem nota reglurnar eru háð því tungumáli sem reglurnar voru gerðar fyrir
  - **gagnaaðferðir** (e. *data-driven methods*) – notað er textasafn sem hefur verið markað og mörkunin yfirfarin handvirkt þannig að hún sé eins rétt og kostur er – forrit er látið læra af gögnunum á tiltekinn hátt, óháð tungumáli.
- Efniviður við gerð markara fyrir íslensku: Textasafn *Íslenskrar orðtíðnibókar* (Jörgen Pind, Stefán Briem og Friðrik Magnússon 1991), notuð bæði við þjálfun gagnamarkara og gerð reglumarkara

Gagnamarkarar

Dæmi um aðferðir og gagnamarkara sem hafa verið prófaðir á íslenskum texta:

- Tölfræðilegar aðferðir
  - Falin Markovslíkön (e. *Hidden Markov Models*): **TnT**
  - Hámarksóreiðuaðferð (e. *maximum entropy*): **MXPOST** (MXP)
- Leiðréttingaaðferð (e. *transformation-based learning*):
  - **µ-TBL** og **fnTBL** (TBL)
- Minnisaðferð (e. *memory-based learning*): **MBT**
- Útkoma úr þjálfun gagnamarkara er þjálfunarlíkan sem má nota við mörkun nýs texta

Reglumarkarar

Hrafn Loftsson bjó til reglumarkarann **IceTagger**

Mörkun íslensks texta

Sigrún Helgadóttir, Hrafn Loftsson og Drezde og Wallenberg hafa gert tilraunir með mörkun íslensks texta. Helstu niðurstöður:

Stakir/samþættir markarar	Nákvæmni
TBL	89,33%
TnT (HMM)	90,54%
IceTagger	91,59%
Tvíátta flokkunar-aðferð	92,06%
HMM+IceTagger	92,19%

Sameinað úttak með kosningu	Nákvæmni
MXP+TBL+TNT	91,45%
TBL+TNT+IceTagger	92,61%
MXP+MBT+TBL+TnT+IceTagger	92,80%
MXP+MBT+TBL*+TnT*+IceTagger*	93,34%

* Nýtir virkni annars markara

Að auki hefur tekist að ná hærri nákvæmni með sjálfstæðu orðsafni, einföldun markaskrár, beitingu málfræðireglna og leiðréttingu á ÍO

Frekari endurbætur á mörkun (HL og fl. 2009):

- Lækka hlutfall óþekktra orða með því að nota stórt orðasafn byggt á BÍN (tilraunir SH og HL lofa góðu)
- Búa til markara (og þáttara) sem byggist á hömlumálfræði (constraint grammar)
- Gera frekari tilraunir með að sameina gagnaaðferðir og reglubyggðar aðferðir
- Finna betri aðferðir til þess að greina fallstjórn sagna

Til hvers er mörkun notuð?

- Tíðnikönnun á texta
- Fyrsta skrefið í:
  - greiningu texta í setningahluta
  - orðtöku úr texta fyrir gerð orðasafns
  - upplýsingaheimt, talkennsl, talgervingu
  - vélrænar þýðingar, orðabókargerð, fyrirspurnarkerfi og gerð leiðréttingarforrita
- Leit í texta verður markvissari, mögulegt verður að finna dæmi um tilteknar setningagerðir

Hvað þarf mörkun að vera nákvæm?

- Í ensku næst yfir 97% nákvæmni í mörkun (markamengi Brown Corpus 87 mörk, markamengi Penn Treebank 45 mörk).
- Íslenska markamengið (ÍO) hefur um 700 mörk og því óraunhæft að ná þessari nákvæmni, núverandi „state-of-the-art nákvæmni er 92,19%
- Í mörgum tilvikum er ekki nauðsynlegt að allir stafir í marki séu réttir
- Nákvæmni í mörkun orðflokks: TnT 98,14%; IceTagger: 97,81%
- Einföldun markamengis: greina ekki sérnöfn, greina ekki milli mismunandi fornafna, greina ekki fallstjórn forsetninga o.s.frv., þá verður nákvæmni hærri (93,55% HMM+Ice)

Mörkuð íslensk málheild

- Verkið er unnið skv. samningi við menntamálaráðuneytið sem fjármagnaði verkefnið
- Stefnt að 25.000.000 lesmálsorðum sem komi úr 900–1.000 textabútum

Hvað er mörkuð málheild (e. *tagged corpus*)?

- Safn fjölbreyttra tölvutækra texta sem hafa verið greindir á málfræðilegan hátt
- Hverjum texta fylgja upplýsingar um textann
- Hverri orðmynd fylgja þær málfræðilegu upplýsingar sem málheildin á að geyma (hér *mark* og *nefnimynd*)
- Textarnir eru skráðir með stöðluðu sniði (XML-útgáfa af TEI-sniði (*Text Encoding Initiative)* fyrir málheildir

Brot úr haus í xml-sniði fyrir skáldsöguna *Mín káta angist* eftir Guðmund Andra Thorsson

```
<monogr>
<title>Mín káta angist</title>
<author born="1957">Guðmundur Andri Thorsson</author>
<imprint>
<publisher>Mál og menning</publisher>
<pubPlace>Reykjavík</pubPlace>
<date value="1988">1988</date>
</imprint>
</monogr>
```

Upphaf skáldsögunnar *Mín káta angist* eftir Guðmund Andra Thorsson, orðmyndir, mörk, nefnimyndir í XML-sniði

```
<body>
<div1>
<p>
<s n="1">
<w type="fp1en" lemma="ég">Ég</w>
<w type="sfg1eþ" lemma="stökkva">stökk</w>
<w type="aa" lemma="á">á</w>
<w type="aþ" lemma="eftir">eftir</w>
<w type="nkeþ" lemma="stræto">stræto</w>
<w type="c" lemma="og">og</w>
<w type="sfg1eþ" lemma="veifa">veifaði</w>
```

- Hverjir nota málheildina?
  - þeir sem vinna að orðabókagerð, margvíslegum máltækniverkefnum og rannsóknum á íslensku nútímamáli
- Að hverju leita þeir?

  t.d. upplýsingum um
  - tíðni orðflokka, orða og beygingarmynda
  - orðasambönd, setningargerð og merkingu
  - hvernig tungumálið er notað á tilteknum tíma, vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti
  - breytileika í máli eftir eiginleikum skrifara, umfjöllunarefni o.s.frv.

- Málheildin er undirstaða fyrir:
  - nútíma orðabókagerð
  - þróun þýðingaforrita
  - þróun máltæknitóla, t.d. fyrir talgreiningu og talgervingu
  - þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði

- Mörg máltæknitól nýtast sérstaklega fyrir blinda (t.d vefþulan Ragga), heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika (Ragga aftur)

Textar í málheildinni

- Textar ritaðir á íslensku 2000 og seinna (2000-2009)
- Upphaflegar hugmyndir (byggðar á BNC, British National Corpus)
- Uppruni
  - 60% úr bókum
  - 25% úr blöðum og tímaritum
  - 5–10% úr öðru útgefnu efni (bæklingar, auglýsingapésar o.s.frv.)
  - 5–10% úr óútgefnu efni (persónuleg bréf, dagbækur, ritgerðir, minnisblöð)
  - <5% úr efni ætluðu til upplestrar (pólitískar ræður, handrit til upplestrar í útvarpi, stólræður o.s.frv.)
  - talmál (Ístal, þingræður, ...)
- Efnisval
  - 25% skáldverk
  - 75% nytjatexti
    - hagnýtt vísindi, náttúrufræði, þjóðfélagsfræði, heimsmál, viðskipti, listir, trúarbrögð, heimspeki, tómstundir,...

Textaöflun

- Ekki er greitt fyrir afnot af textum
- Engir textar notaðir án leyfis rétthafa
- Stuðningsyfirlýsingar frá Rithöfundasambandi Íslands, Hagþenki og Félagi íslenskra bókaútgefenda

Hvar má finna efnið?

- Beint frá útgefendum (bækur, blöð og tímarit)
- Af vefsíðum (vefblöð, veftímarit, blogg, tölvupóstur, greinar, lög, reglugerðir o.s.frv.)
- Beint frá rétthöfum (skólaritgerðir, bækur)

Vinnulag við efnisöflun og úrvinnslu

- Haft samband við rétthafa og beðið um leyfi (um síma, um tölvupóst, með bréfi í pósti)
- Texti sóttur (til útgefanda sem er ekki alltaf rétthafi: pdf, Word, hreinn texti, annað)
- Texti dreginn úr umbroti, aukaefni eytt (neðanmálsgreinum, myndum töflum, efnisyfirliti, heimildaskrám, erlendum bútum, löngum tilvitnunum í eldri rit), ýmis tákn löguð, stafatafla samræmd (allt efni sett í UTF-8), 20% skorin af textum úr útgefnum bókum
- Bókfræðilegar upplýsingar skráðar (titill, höfundur, útgefandi, útgáfuár, ýmis flokkun)

Staða efnisöflunar í mars 2009 (áætlun)	þús. orða	%
Blogg	1.964	10,2
Bækur	2.018	10,5
Dagblöð, prentuð og á vef	7.376	38,3
Til upplestrar	488	2,5
Lög, reglug., dómar	2.132	11,1
Skólaritgerðir	722	3,7
Ýmsir bæklingar	33	0,2
Textavarp	47	0,2
Tímarit, prentuð og á vef	1.952	10,1
Tölvupóstlistar	121	0,6
Vefsetur	1.069	5,6
Vísindavefur	1.336	6,9
Alls	19.257	100,0

Mörkun og önnur úrvinnsla

- Tilreiðsla (tokenization, texta skipt í orð með sérstöku forriti)
- Málfræðileg mörkun (IceTagger eða samsettur markari)
- Lemmun (Lemmald, lemmari Antons)
- Texti færður í xml-snið (TEI, Text Encoding Initiative)
- Texta komið fyrir í gagnasafni

Hvernig verður aðgangur að málheildinni?

- Allt efni sem er safnað á vegum málheildar er líka notað til þess að styrkja Textasafn OH og verður þar aðgengilegt í opnum aðgangi
  - Takmarkaður uppflettiaðgangur á vefsetri SÁ (t.d. 50–300 orð á undan og eftir orði sem leitað er að). Textar úr Morgunblaðinu, af Vísindavef, bloggtextar og textar nokkurra skáldsagna og ævisagna eru þegar aðgengilegir í Textasafninu
- Textar í Markaðri íslenskri málheild
  - Leitaraðgangur á vefsetri OH með sérstöku leitarforriti, leita eftir textaflokki o.fl., mörkun nýtt
  - Dreifing á geisladiskum eða með því að sækja skrár á vefsetur SÁ. Notendur undirrita notkunarleyfi eða samþykkja notkunarskilmála á vefsetri

Takk fyrir áheyrnina!

sigruhel@hi.is

orð	nefnimynd	mark	skýring
ég	ég	fp1en	f: fn; p: pfn; 1: 1. pers.; e: et.; n: nefnifall
stökk	stökkva	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
á	á	aa	a: ao.; a: stýrir ekki falli
eftir	eftir	aþ	a: ao.; þ: stýrir þágufalli
stræto	stræto	nkeþ	n: no.; k: kk.; e: et.; þ: þgf.
og	og	c	c: samtenging
veifaði	veifa	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
,	,	,	komma
vagnstjórinn	vagnstjóri	nkeng	n: no.; k: kk.; e: et.; n: nf.; g: með greini
sá	sjá	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
mig	ég	fp1eo	f: fn.; p: pfn.; 1: 1. pers.; e: et.; n: þolfall
og	og	c	c: samtenging
stoppaði	stoppa	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
.	.	.	punktur

Mynd 1. Greining orða í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Textasafn *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem. 1991. Orðabók Háskólans.)

100 textar, um 5000 orð hver (590.000 lesmálsorð)

frumsamin íslensk skáldverk	20 %
þýdd skáldverk	20 %
ævisögur	20 %
nytjatexti	20 % (hugvísindi og raunvísindi)
barnabækur	20 % (frumsamdar og þýddar)

Markamengi um 700 mörk

Til hvers verður málheildin notuð og af hverjum?

Notendur: þeir sem vinna að orðabókagerð, margvíslegum máltækniverkefnum og rannsóknum á íslensku nútímamáli.

Úr málheildinni má lesa ýmiss konar gagnlegan fróðleik, t.d.

- um tíðni orðflokka, orða og beygingarmynd,
- um orðasambönd, setningargerð og merkingu
- um hvernig tungumálið er notað á tilteknum tíma, vísbendingar um orðaforðann og einnig um málfræðilega og setningarfræðilega þætti
- um breytileika í máli eftir eiginleikum skrifara, umfjöllunarefni o.s.frv.

Undirstaða fyrir:

- þróun þýðingarforrita
- nútíma orðabókagerð
- þróun máltæknitóla, t.d. fyrir talgreiningu og talgervingu
- þróun hjálparforrita með ritvinnslu, t.d. forrita sem leiðbeina um stafsetningu og málfræði.

Mörg máltæknitól nýtast sérstaklega fyrir blinda (t.d vefþulan Ragga), heyrnarskerta og hreyfihamlaða og einnig þá sem glíma við skriftar- og lestrarörðugleika (Ragga aftur)

Mörkun íslensks texta

- Sigrún Helgadóttir gerði tilraunir með gagnamarkara (námfúsa markara) 2002–2004, prófaðir 4 gagnamarkar og fleiri aðferðir, t.d. sameining markara, notkun orðasafns, einföldun markaskrár, beiting málfræðireglna
  - Besti markarinn, TnT, gaf 90,4% nákvæmni, besta niðurstaða 93,7% fékkst með samsettri aðferð (kosningu milli markara, notkun orðasafns, beitingu málfræðireglna, einföldun markaskrár)
- Hrafn Loftsson bjó til reglubyggðan markara, IceTaggger, (2006?) og náði um 91,6% nákvæmni
  - Með endurbótum og samsettum aðferðum hefur náðst **92,19**% nákvæmni þar sem IceTagger er grunntól
  - Með einföldun markaskrár og með því að leiðrétta villur í málheild ÍO hefur náðst 93,55% nákvæmni
- Dredze og Wallenberg (2008) byggðu á þeirri vinnu sem þegar hafði verið unnin, beittu m.a. tvíátta flokkunaraðferð og náðu 92,06% nákvæmni