

# T-(538|725)-MALV, Málvinnsla Uppbygging og nýnæmi upplýsinga

Hrafn Loftsson<sup>1</sup> Hannes Högni Vilhjálmsson<sup>1</sup>

<sup>1</sup>Tölvunarfræðideild, Háskólinn í Reykjavík

Október 2007

**1** Orðræðugreining (örstutt upprifjun)

**2** Nýnæmi upplýsinga

**3** Uppbygging upplýsinga

**1** Orðræðugreining (örstutt upprifjun)

**2** Nýnæmi upplýsinga

**3** Uppbygging upplýsinga

# Orðræðugreining (upprifjun)

## Markmið orðræðugreiningar

Finna mynstur í venjulegri málnotkun fólks sem lýsir því hvernig **orðræðuáhöld** (e. Discourse Devices) eru notuð til að ná ýmsum **orðræðumarkmiðum** (e. Discourse Functions/Intentions).

## Vél getur nýtt þetta til að...

- ...túlka, skilja og vinna úr mennskum texta,
- ...eiga í samskiptum við fólk á þeirra forsendum,
- ...aðstoða fólk við samskipti þeirra á milli.

# Orðræðugreining (upprifjun)

## Samhengi og orðræðulíkan

**Orðræðulíkan** (e. Discourse Model) heldur utanum **Samhengi** textans (e. Context) og gegnir lykilhlutverki í réttri túlkun orðræðunnar. Í því eru meðal annars þær upplýsingar sem hafa komið fram í textanum, á formi **orðræðueininga** (e. Discourse Entities).

1 Orðræðugreining (örstutt upprifjun)

2 Nýnæmi upplýsinga

3 Uppbygging upplýsinga

## Upplýsingaástand (e. Information Status)

Hvað gerir framleiðandi ráð fyrir að viðtakandi viti mikið um tilteknar upplýsingar þegar þær eru bornar fram?

- "Gamlar" / "Gefnar" (e. Old, Given) upplýsingar.  
Samsvarandi einingar eru tiltækar í orðræðulíkani.
- "Ályktaðar" / "Miðlaðar" (e. Inferred, Mediated) upplýsingar.  
Samsvarandi einingar eru óbeint tiltækar.
- "Nýjar" (e. New) upplýsingar.  
Samsvarandi einingar eru ekki tiltækar.

## Gamlar upplýsingar

**Gamlar** upplýsingar felast í öllum beinum vísunum á orðræðueiningar sem þegar eru í orðræðulíkaninu. T.d. þegar ákveðinn greinir eða eignarfornöfn eru notuð.

## Dæmi

A1: "Ég sá kött í gær"

A2: "**Kötturinn** var bröndóttur"



## Ályktaðar upplýsingar

Þótt ekki sé beinlínis vísað á einingu sem til er í líkaninu, þá getur verið um upplýsingar að ræða sem viðtakandinn getur auðveldlega **ályktað** útfrá þeim einingum sem þar eru eða útfrá almennri þekkingu. Dæmi um nokkrar tegundir ályktaðra upplýsinga:

- "Sólin sést ekki lengur" (almenn þekking)
- "Þegar ég kom heim beið hundurinn fyrir innan **dyrnar**" (þekktur hluti af stærri hlut)
- "Í vísindaferðinni var ekki nóg af **bjór**" (þekkt atvik)
- "Ég drep **peðið** með **riddaranum**" (þekkt umhverfi)

## Nýjar upplýsingar

**Nýjar** upplýsingar eru allar upplýsingar sem ekki eru gamlar eða ályktaðar.

## Dæmi

"Ég held ég hafi séð **bleika geimveru!**"

1 Orðræðugreining (örstutt upprifjun)

2 Nýnæmi upplýsinga

3 Uppbygging upplýsinga

## Grice heilræðin um skýrleika og upplýsingamagn

Við þurfum að finna jafnvægi milli **skýrleika** og **upplýsingamagns** þegar við búum til nýjar setningar.

- Skýrt mál krefst góðrar **samfellu** í máli (e. Coherence). Setningar þurfa að tengjast og það krefst svolítillrar endurtekningar.
- Áhersla á gott upplýsingamagn gefur ekki mikið svigrúm til þess að halda í það sem búið er að segja.

## Uppbygging upplýsinga (e. Information Structure)

Þegar framleiðendur pakka upplýsingum inn í setningar, þá þurfa þeir bæði að koma fyrir gömlum upplýsingum, sem tengja setninguna við það sem komið er á undan, og nýjum upplýsingum sem halda samskiptunum lifandi.

Þannig má skipta setningum í tvo hluta:

- **Pema** (e. Theme, Topic)  
Tenging setningar við fyrri setningar. Um hvað er fjallað?
- **Rema** (e. Rheme, Comment)  
Nýjar upplýsingar. Hvað er sagt um það sem fjallað er um?

## Dæmi

A1: "[Ég **P**] [sá góða mynd í gær **R**]."

A2: "[Hún fjallaði um **P**] [heilbrigðiskerfið í Bandaríkjunum **R**]"

## Orðræðulíkanið uppfært

Í þessu dæmi sést vel að orðræðueiningarnar sem vísað er í eru þekktar í þema hlutunum en búa verður til nýjar einingar í remu hlutunum.

Orðræðueiningarnar samsvara þarna mögulegum **anafórum** (e. Anaphors) og **undanförum** (e. Antecedents), eða því sem einnig eru kallaðar **afturleitar miðjur** (e. Backward-looking Centers) og **framleitar miðjur** (e. Forward-looking Centers).

## Dæmi

A1: "[Ég **P**] [sá góða mynd í gær **R**]."

A2: "[Hún fjallaði um **P**] [heilbrigðiskerfið í Bandaríkjunum **R**]"

## Orðræðulíkanið uppfært

Í þessu dæmi sést vel að orðræðueiningarnar sem vísað er í eru þekktar í þema hlutunum en búa verður til nýjar einingar í remu hlutunum.

Orðræðueiningarnar samsvara þarna mögulegum **anafórum** (e. Anaphors) og **undanförum** (e. Antecedents), eða því sem einnig eru kallaðar **afturleitar miðjur** (e. Backward-looking Centers) og **framleitar miðjur** (e. Forward-looking Centers).

## Samsvarar ekki frumlagi/andlagi

Athugið að setning sem hefur eina tiltekna málfræðilega greiningu, og þar af leiðandi ákveðið frumlag og andlag, getur haft mismunandi upplýsingagreiningu eftir samhengi (þ.e. eftir því hvað er komið í orðræðulíkanið).

## Dæmi

A: "Hver stal veskinu?" B: "[Jón R] [stal veskinu P]"

A: "Hverju stal Jón?" B: "[Jón stal P] [veskinu R]"



## Samsvarar ekki frumlagi/andlagi

Athugið að setning sem hefur eina tiltekna málfræðilega greiningu, og þar af leiðandi ákveðið frumlag og andlag, getur haft mismunandi upplýsingagreiningu eftir samhengi (þ.e. eftir því hvað er komið í orðræðulíkanið).

## Dæmi

A: "Hver stal veskinu?" B: "[Jón **R**] [stal veskinu **P**]"

A: "Hverju stal Jón?" B: "[Jón stal **P**] [veskinu **R**]"

## Nýnæmið gefið í skyn

Framleiðendur beita oft orðræðuáhöldum til að auðvelda viðtakendum að finna hið nýja í texta. Þessi áhöld eru breytileg eftir tungumálum.

## Dæmi (áhald=orðaröð)

A: "Hver borðaði kökuna?" B: "Anna borðaði kökuna"

A: "Hvað borðaði Anna?" B: "Kökuna borðaði Anna"

## Dæmi (áhald=raddítónun)

A: "Hver er hér?" B: "**Batman** er hér"

A: "Hvar er Batman?" B: "Batman er **hér**"

# Uppbygging upplýsinga

## Nýnæmið gefið í skyn

Framleiðendur beita oft orðræðuáhöldum til að auðvelda viðtakendum að finna hið nýja í texta. Þessi áhöld eru breytileg eftir tungumálum.

## Dæmi (áhald=orðaröð)

A: "Hver borðaði kökuna?" B: "Anna borðaði kökuna"

A: "Hvað borðaði Anna?" B: "Kökuna borðaði Anna"

## Dæmi (áhald=raddítónun)

A: "Hver er hér?" B: "**Batman** er hér"

A: "Hvar er Batman?" B: "Batman er **hér**"

## Nýnæmið gefið í skyn

Framleiðendur beita oft orðræðuáhöldum til að auðvelda viðtakendum að finna hið nýja í texta. Þessi áhöld eru breytileg eftir tungumálum.

## Dæmi (áhald=orðaröð)

A: "Hver borðaði kökuna?" B: "Anna borðaði kökuna"

A: "Hvað borðaði Anna?" B: "Kökuna borðaði Anna"

## Dæmi (áhald=raddítónun)

A: "Hver er hér?" B: "**Batman** er hér"

A: "Hvar er Batman?" B: "Batman er **hér**"

## Fókus (e. Focus)

- Innan þema og rema má gera ákveðnar upplýsingar meira áberandi (e. Salient) með því að gefa þeim sérstakan fókus.
- Þegar gamlar upplýsingar eru gerðar áberandi eða nýjar upplýsingar sérstaklega áberandi, þá hefur það ákveðna þýðingu.
- Oftast er um að ræða samanburð.

## Dæmi

A: "Ég á gult hjól, en þú?"

B: "[Ég á **P**] [gulan **bíl R**]!"

## Fókus (e. Focus)

- Innan þema og rema má gera ákveðnar upplýsingar meira áberandi (e. Salient) með því að gefa þeim sérstakan fókus.
- Þegar gamlar upplýsingar eru gerðar áberandi eða nýjar upplýsingar sérstaklega áberandi, þá hefur það ákveðna þýðingu.
- Oftast er um að ræða samanburð.

## Dæmi

A: "Ég á gult hjól, en þú?"

B: "[Ég á **P**] [gulan **bíl R**]!"

## Að lokum...

- Fólk fer létt með að pakka of afpakka upplýsingum á þjálan máta og virðist nær alltaf vera með á hreinu hvað er gefið og hvað er nýtt. Þetta er aftur á móti mjög erfitt fyrir vélar vegna flókins samhengis og þörf á góðri ályktunarhæfni.
- Þegar tölvur lesa texta upphátt heyrir einmitt vel hversu slæmur lesskilningur þeirra er. Þær leggja oft áherslu á upplýsingar sem við sjáum sem gamlar eða ályktaðar, og gleyma svo að leggja áherslu á nýjar upplýsingar.
- Tölvutal er því sú grein innan tölvunarfræðinnar sem hefur hvað mest nýtt sér rannsóknir á þessu sviði málvinnslu.