

Kvantitative metoder inden for korpuslingvistiske projekter

*Illustreret ved eksempler fra Den Danske Ordbogs
korpus, Korpus 2000 og Korpus 90*

Jørg Asmussen

Det Danske Sprog- og Litteraturselskab, DSL

ja@dsl.dk

1. Grundlæggende metodiske forhold
2. Korpuslingvistik – en introduktion
3. Finde, tælle, konkludere
4. Sammenligninger – hvordan?
5. Struktur og kollokation
6. Heuristisk tekstklassifikation
7. Litteratur m.m.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Grundlæggende metodiske forhold

Øvelse 1.1 (15 minutter)

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed		
Forskerens position i forhold til disse fænomener		
Måden, fænomenerne kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	objekter	subjekter
Forskerens position i forhold til disse fænomener		
Måden, fænomenerne kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	objekter	subjekter
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	objekter	subjekter
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk, heuristisk	hermeneutisk
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	objekter	subjekter
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk, heuristisk	hermeneutisk
Præsentation af resultater	numerisk	tekst
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	objekter	subjekter
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk, heuristisk	hermeneutisk
Præsentation af resultater	numerisk	tekst
Afprøvning af hypoteser	bevis	begrundelse

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	intersubjektivitet	
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk, heuristisk	hermeneutisk
Præsentation af resultater	numerisk	tekst
Afprøvning af hypoteser	bevis	begrundelse

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	intersubjektivitet	
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk, heuristisk	hermeneutisk
Præsentation af resultater	numerisk	tekst
Afprøvning af hypoteser	bevis	begrundelse

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	intersubjektivitet	
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk heuristisk	kritisk
Præsentation af resultater	numerisk	tekst
Afprøvning af hypoteser	bevis	begrundelse

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...

	Kvantitativ	Kvalitativ
Holdningen til fænomenernes grundlæggende beskaffenhed	intersubjektivitet	
Forskerens position i forhold til disse fænomener	observatør	deltager
Måden, fænomenerne kategoriseres på	statistisk heuristisk	kritisk
Præsentation af resultater	kodifikation	
Afprøvning af hypoteser	bevis	begrundelse

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

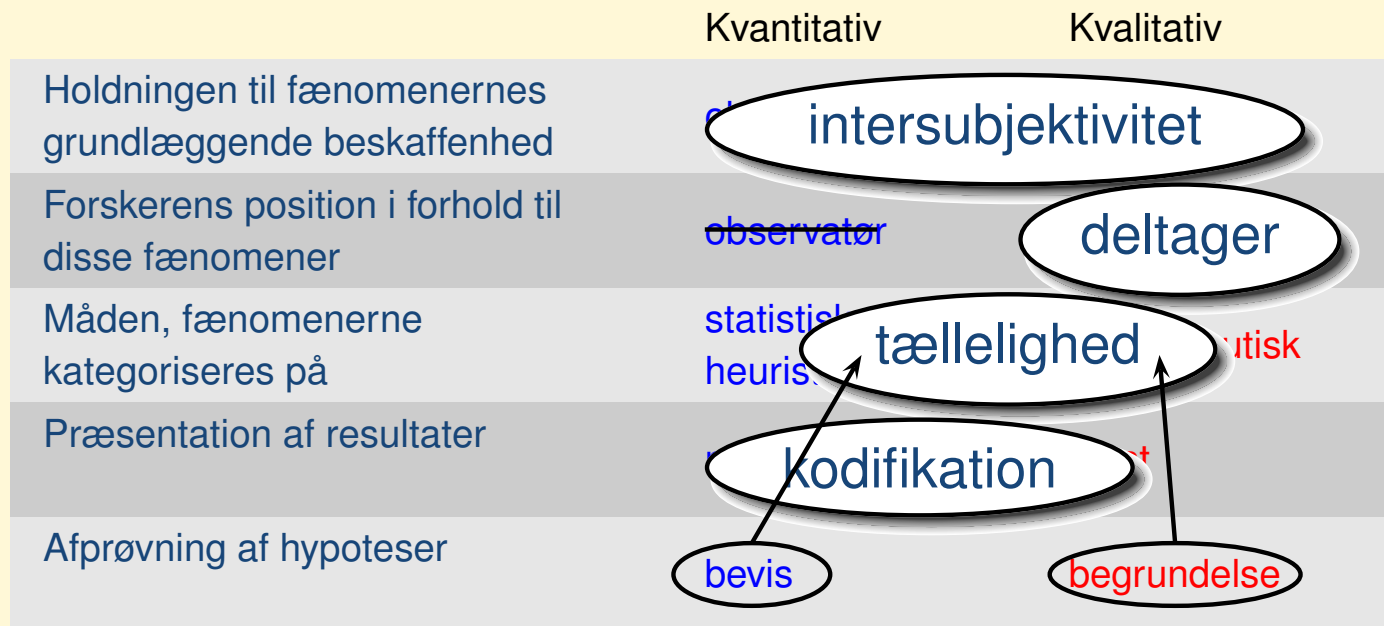
Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Kvantitativ vs. kvalitativ

Overvej, hvorved den kvantitative metode adskiller sig fra den kvalitative mht. ...



Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed		
Forskerens position i forhold til sproget		
Måden, sproglige fænomener kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	fuzzy	regelbundet
Forskerens position i forhold til sproget		
Måden, sproglige fænomener kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	fuzzy	regelbundet
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på		
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	fuzzy	regelbundet
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statistik	introspektion
Præsentation af resultater		
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	fuzzy	regelbundet
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statistik	introspektion
Præsentation af resultater	numerisk	binære regler
Afprøvning af hypoteser		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	fuzzy	regelbundet
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statistik	introspektion
Præsentation af resultater	numerisk	binære regler
Afprøvning af hypoteser	empirisk	rationalistisk

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	f	regelmæssighed
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statistik	introspektion
Præsentation af resultater	numerisk	binære regler
Afprøvning af hypoteser	empirisk	rationalistisk

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	f	regelmæssighed
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statistik	introspektion
Præsentation af resultater	numerisk	binære regler
Afprøvning af hypoteser	empirisk	rationalistisk

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	f	regelmæssighed
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statist	tællelighed
Præsentation af resultater	numerisk	binære regler
Afprøvning af hypoteser	empirisk	rationalistisk

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...

	Probabilitet	Grammatikalitet
Holdningen til sprogets grundlæggende beskaffenhed	f	regelmæssighed
Forskerens position i forhold til sproget	observatør	deltager
Måden, sproglige fænomener kategoriseres på	statist	tællelighed
Præsentation af resultater	kodifikation	regler
Afprøvning af hypoteser	empirisk	rationalistisk

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

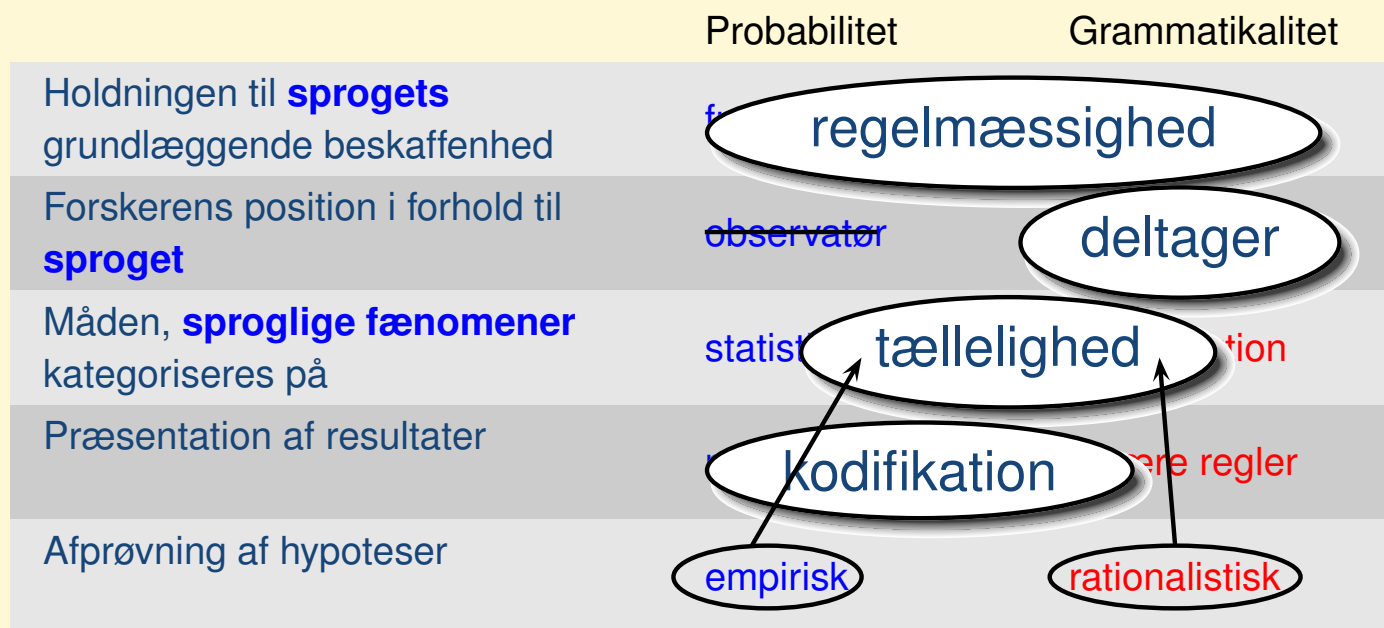
Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Probabilitet vs. grammatikalitet

Dikotomien *kvantitativ* vs. *kvalitativ* kommer i (strukturel) lingvistik bl.a. til udtryk ved ...



Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse		
Materialegrundlaget		
Metode		
Beskrivelse af sproglig regelmæssighed		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget		
Metode		
Beskrivelse af sproglig regelmæssighed		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode		
Beskrivelse af sproglig regelmæssighed		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved . . .

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed		

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved . . .

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed	sandsynligheder	regler

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed	sandsynligheder	regler

Altså er *I live in New York*
mere grammatisk end *I live
in Dayton, Ohio* ...

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed	sandsynligheder	regler

Altså er *I live in New York*
mere grammatisk end *I live in Dayton, Ohio* ...

One doesn't study all of botany by making artificial flowers

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed	sandsynligheder	regler

Grundantagelsen her: målet er at beskrive og måske forklare sprogets regelmæssigheder. Ingen teoretisk ramme eller metode klarer dette fuldkomment, men belyser forskellige sproglige aspekter fra forskellige vinkler. Ingen er *per se* bedre end de andre, men alle kan supplere hinanden.

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Empirisme vs. rationalisme

Inden for sprogbeskrivelsen er den empiriske over for den rationalistiske tilgang bl.a. karakteriseret ved ...

	Empirisme	Rationalisme
Genstanden for beskrivelse	performans	kompetens
Materialegrundlaget	tekstkorpora	sproglige ytringer
Metode	optælling	introspektion
Beskrivelse af sproglig regelmæssighed	sandsynligheder	regler

Program

Grundlæggende

Kvantitativ vs. kvalitativ

Probabilitet vs. grammatikalitet

Empirisme vs. rationalisme

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

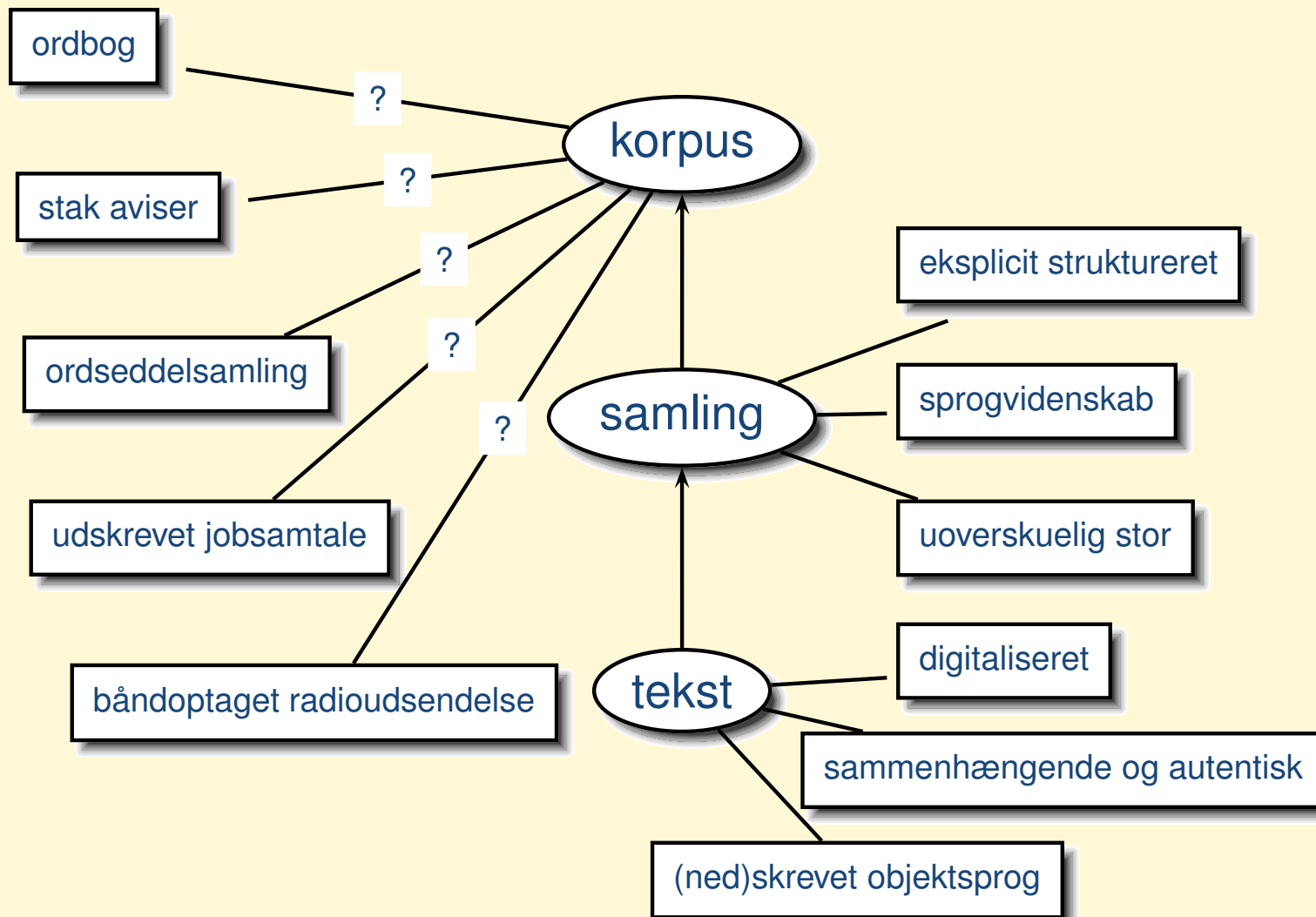
Litteratur m.m.

Grundantagelsen her: målet er at beskrive og måske forklare sprogets regelmæssigheder. Ingen teoretisk ramme eller metode klarer dette fuldkomment, men belyser forskellige sproglige aspekter fra forskellige vinkler. Ingen er *per se* bedre end de andre, men alle kan supplere hinanden.

Korpuslingvistik – en introduktion

Opgave 2.1 (10 minutter)

Definition: korpus



Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

■ Den Danske Ordbogs Korpus (DDOC):

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på tekstniveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på tekstniveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk
 - ◆ særlig: download og ekspert-interface

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk
 - ◆ særlig: download og ekspert-interface
- Korpus 2000 (K2000):

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk
 - ◆ særlig: download og ekspert-interface
- Korpus 2000 (K2000):
 - ◆ 28 mio. ords løbende tekst

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på tekstniveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk
 - ◆ særlig: download og ekspert-interface
- Korpus 2000 (K2000):
 - ◆ 28 mio. ords løbende tekst
 - ◆ skriftsprog, 1998 til 2002

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Referencekorpora for dansk

- Den Danske Ordbogs Korpus (DDOC):
 - ◆ 40 mio. ords løbende tekst
 - ◆ skrift- og talesprog, 1983-1992
 - ◆ kommunikativt og sociolingvistisk inspireret, på teksthiveau
 - ◆ download fra www.korpus2000.dk sammen med konkordansværktøjet Semaskop
- Korpus 90 (K90):
 - ◆ 28 mio. ords løbende tekst, delmængde af DDOC
 - ◆ skriftsprog, 1983-1992
 - ◆ lemma, morfologi og syntaks, på ordniveau
 - ◆ online tilgængelig på www.korpus2000.dk
 - ◆ særlig: download og ekspert-interface
- Korpus 2000 (K2000):
 - ◆ 28 mio. ords løbende tekst
 - ◆ skriftsprog, 1998 til 2002
 - ◆ svarende til K90 → sammenlignende undersøgelser

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som ‘modsætning’ til *introspektiv lingvistik*

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som 'modsætning' til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som 'modsætning' til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som 'modsætning' til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som 'modsætning' til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som 'modsætning' til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.
- Årsag til fokusændringen: it-udviklingen i de seneste 20 år

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som ‘modsætning’ til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.
- Årsag til fokusændringen: it-udviklingen i de seneste 20 år
- Korpuslingvistikken er – som metodevidenskab – i dag

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som ‘modsætning’ til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.
- Årsag til fokusændringen: it-udviklingen i de seneste 20 år
- Korpuslingvistikken er – som metodevidenskab – i dag
 - ◆ stærkt datalingvistisk orienteret

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som ‘modsætning’ til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.
- Årsag til fokusændringen: it-udviklingen i de seneste 20 år
- Korpuslingvistikken er – som metodevidenskab – i dag
 - ◆ stærkt datalingvistisk orienteret
 - ◆ tiltagende datalogisk/matematisk orienteret

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Beskrivelse: korpuslingvistik

- Oprindelig vel formet som ‘modsætning’ til *introspektiv lingvistik*
- Med tiden er *lingvistikken* dog blevet nedtonet: sprogbeskrivelse er sprogbeskrivelse uanset materialegrundlaget. . .
- Nu snarere: videnskab om metoder til
 - ◆ *opbygning* af korpora
 - ◆ *udnyttelse* af korpora m.m.
- Årsag til fokusændringen: it-udviklingen i de seneste 20 år
- Korpuslingvistikken er – som metodevidenskab – i dag
 - ◆ stærkt datalingvistisk orienteret
 - ◆ tiltagende datalogisk/matematisk orienteret
 - ◆ . . . og står derfor svagt i Danmark

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

■ Korpuslingvistisk:

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

- Korpuslingvistisk:
 - ◆ som beror på state-of-the-art korpuslingvistisk metodologi

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

- Korpuslingvistisk:
 - ◆ som beror på state-of-the-art korpuslingvistisk metodologi
 - ◆ stikord: algoritmisk-statistisk analyse

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

- Korpuslingvistisk:
 - ◆ som beror på state-of-the-art korpuslingvistisk metodologi
 - ◆ stikord: algoritmisk-statistisk analyse
- Korpusbaseret:

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

- Korpuslingvistisk:
 - ◆ som beror på state-of-the-art korpuslingvistisk metodologi
 - ◆ stikord: algoritmisk-statistisk analyse
- Korpusbaseret:
 - ◆ som beror på et korpus som grundlæggende materiale

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Korpuslingvistisk eller -baseret?

- Korpuslingvistisk:
 - ◆ som beror på state-of-the-art korpuslingvistisk metodologi
 - ◆ stikord: algoritmisk-statistisk analyse
- Korpusbaseret:
 - ◆ som beror på et korpus som grundlæggende materiale
 - ◆ stikord: lingvistens umiddelbare kvantitative og kvalitative iagttagelser

Program

Grundlæggende

Korpuslingvistik

Definition: korpus

Referencekorpora for dansk

Beskrivelse: korpuslingvistik

Korpuslingvistisk eller -baseret?

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

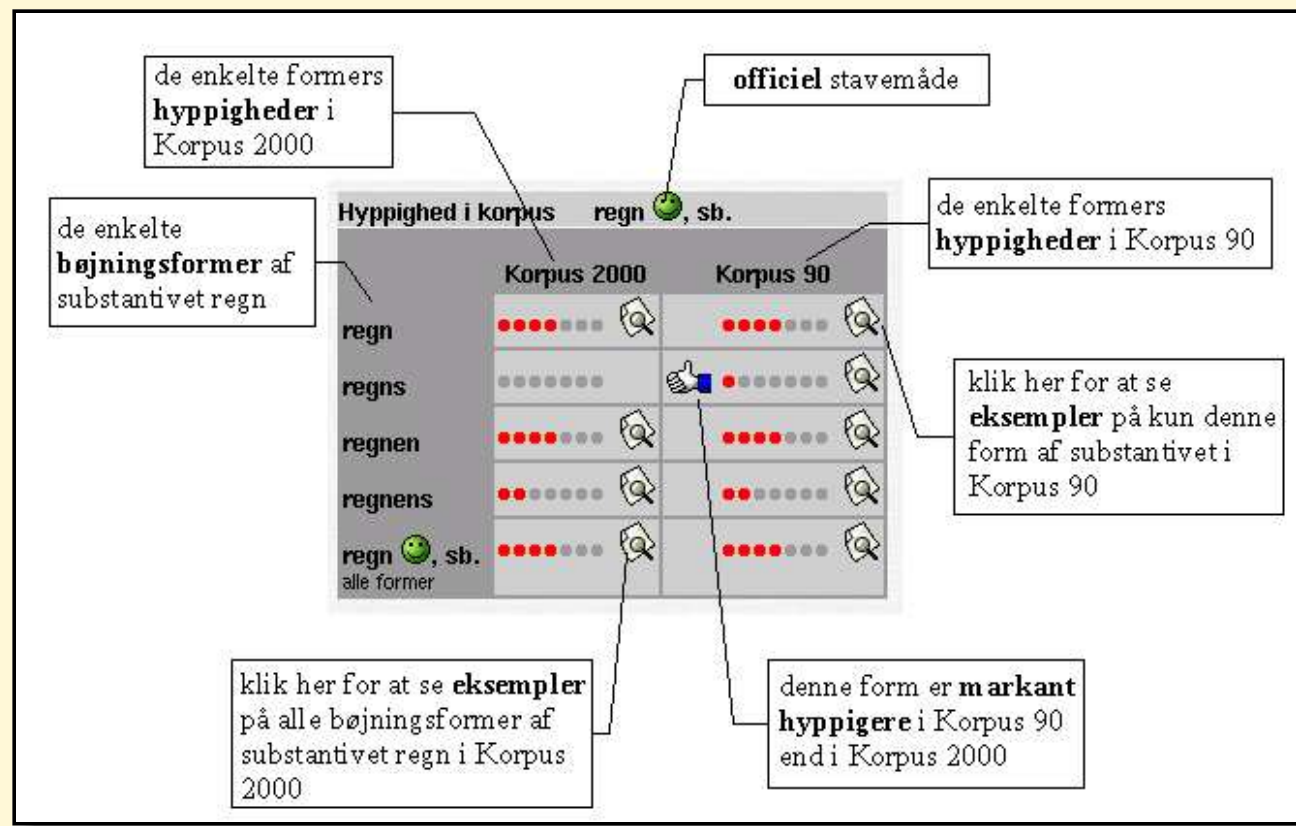
Heuristisk tekstklassifikation

Litteratur m.m.

Finde – tælle – konkludere

Opgave 3.1 (10 minutter)

Eksempel *regn*



Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

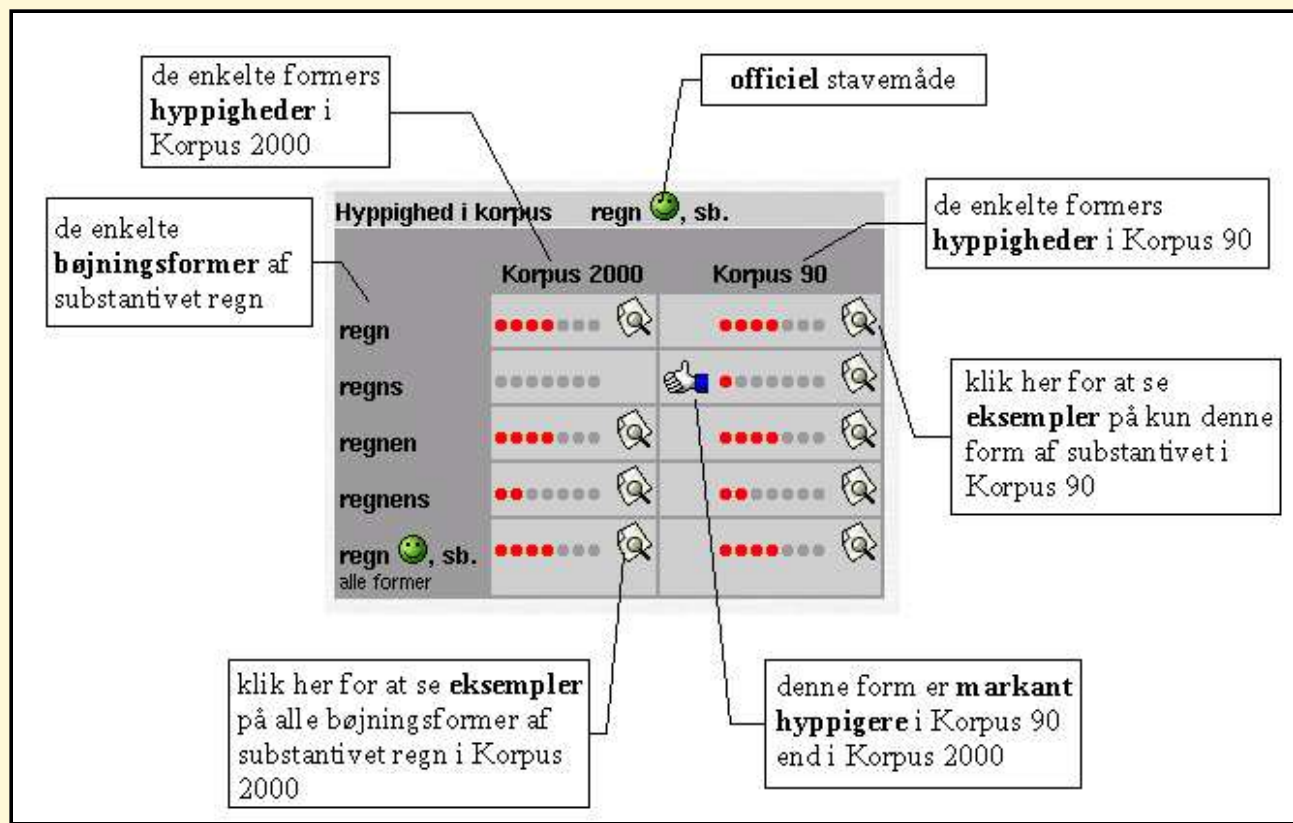
Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *regn*



Konklusion: øjensynlig ingen nævneværdig hyppighedsforskel i de to korpora

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *mobiltelefon* (K2000: ~1500, K90: ~60)

	Korpus 2000	Korpus 90
mobiltelefon	          	          
mobiltelefons	          	          
mobiltelefonen	          	          
mobiltelefonens	          	          
mobiltelefoner	          	          
mobiltelefoners	          	          
mobiltelefonerne	          	          
mobiltelefonernes	          	          
mobiltelefon 🟢, sb.	          	          

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

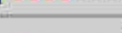
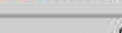
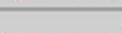
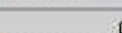
Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *mobiltelefon* (K2000: ~1500, K90: ~60)

	Korpus 2000	Korpus 90
mobiltelefon	  	 
mobiltelefons	  	 
mobiltelefonen	  	 
mobiltelefonens	  	 
mobiltelefoner	  	 
mobiltelefoners	 	 
mobiltelefonerne	  	 
mobiltelefonemes	  	 
mobiltelefon 🟢, sb.	  	 

Konklusion: hyppighedsforskel formentlig betinget af den tekniske udvikling, som afspejler sig i sprogbrugen

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *biltelefon* (K2000: 9, K90: 51)

	Korpus 2000	Korpus 90
biltelefon	●●○○○○○	●●○○○○○
biltelefons	○○○○○○○	○○○○○○○
biltelefonen	●●○○○○○	●●○○○○○
biltelefonens	○○○○○○○	○○○○○○○
biltelefoner	●○○○○○○○	●●○○○○○
biltelefoners	○○○○○○○	○○○○○○○
biltelefoneme	○○○○○○○	●○○○○○○○
biltelefonernes	○○○○○○○	●○○○○○○○
biltelefon 🟢, sb.	●●○○○○○	●●○○○○○

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *biltelefon* (K2000: 9, K90: 51)

	Korpus 2000	Korpus 90
biltelefon	●●●●●●	●●●●●●
biltelefons	●●●●●●	●●●●●●
biltelefonen	●●●●●●	●●●●●●
biltelefonens	●●●●●●	●●●●●●
biltelefoner	●●●●●●	●●●●●●
biltelefoners	●●●●●●	●●●●●●
biltelefoneme	●●●●●●	●●●●●●
biltelefonernes	●●●●●●	●●●●●●
biltelefon ☺, sb.	●●●●●●	●●●●●●

Konklusion: hyppighedsforskel formentlig igen betinget tekniske forandringer, som afspejler sig i sprogbrugen

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *kambrium* (K2000: 0, K90: 4)

	Korpus 2000	Korpus 90
kambrium		   
kambriums		
kambrium 😊, sb.		   

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel *kambrium* (K2000: 0, K90: 4)

	Korpus 2000	Korpus 90
kambrium		👍 🔍
kambriums		
kambrium 😊, sb.		👍 🔍

Konklusion: tilfælde eller uensartet sammensatte korpora?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Morfologi: genitiver af substantiver

Genitivformer af:	Korpus 2000	Korpus 90
bil	393	586
cykel	23	35
mand	1606	1936

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Morfologi: genitiver af substantiver

Genitivformer af:	Korpus 2000	Korpus 90
bil	393	586
cykel	23	35
mand	1606	1936

Genitivformer af:	Korpus 2000	Korpus 90
bil	4,7%	5,7%
cykel	1,7%	2,0%
mand	6,5%	6,5%

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

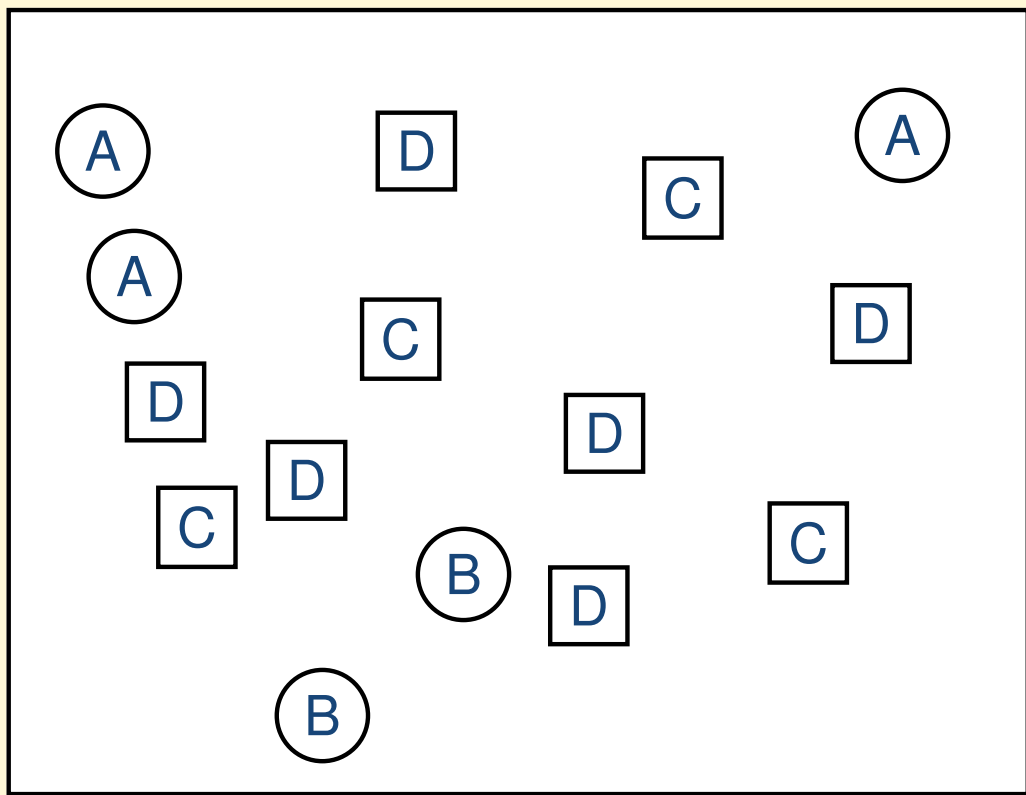
Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Fortolkning af data

Diskuter på baggrund af nedenstående figur udsagnet:
I materialet er der flere forekomster af C end af B.



Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

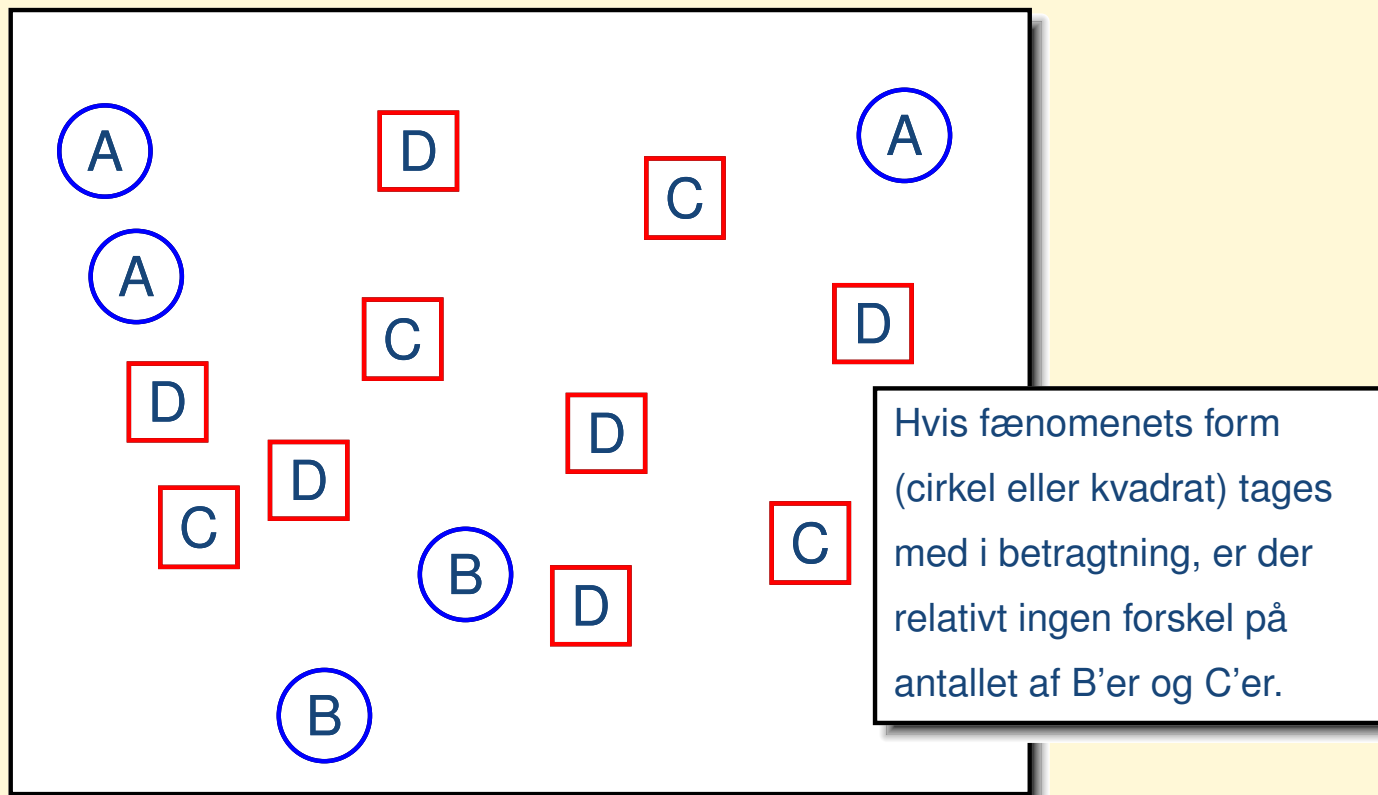
Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Fortolkning af data

Diskuter på baggrund af nedenstående figur udsagnet:
I materialet er der flere forekomster af C end af B.



Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

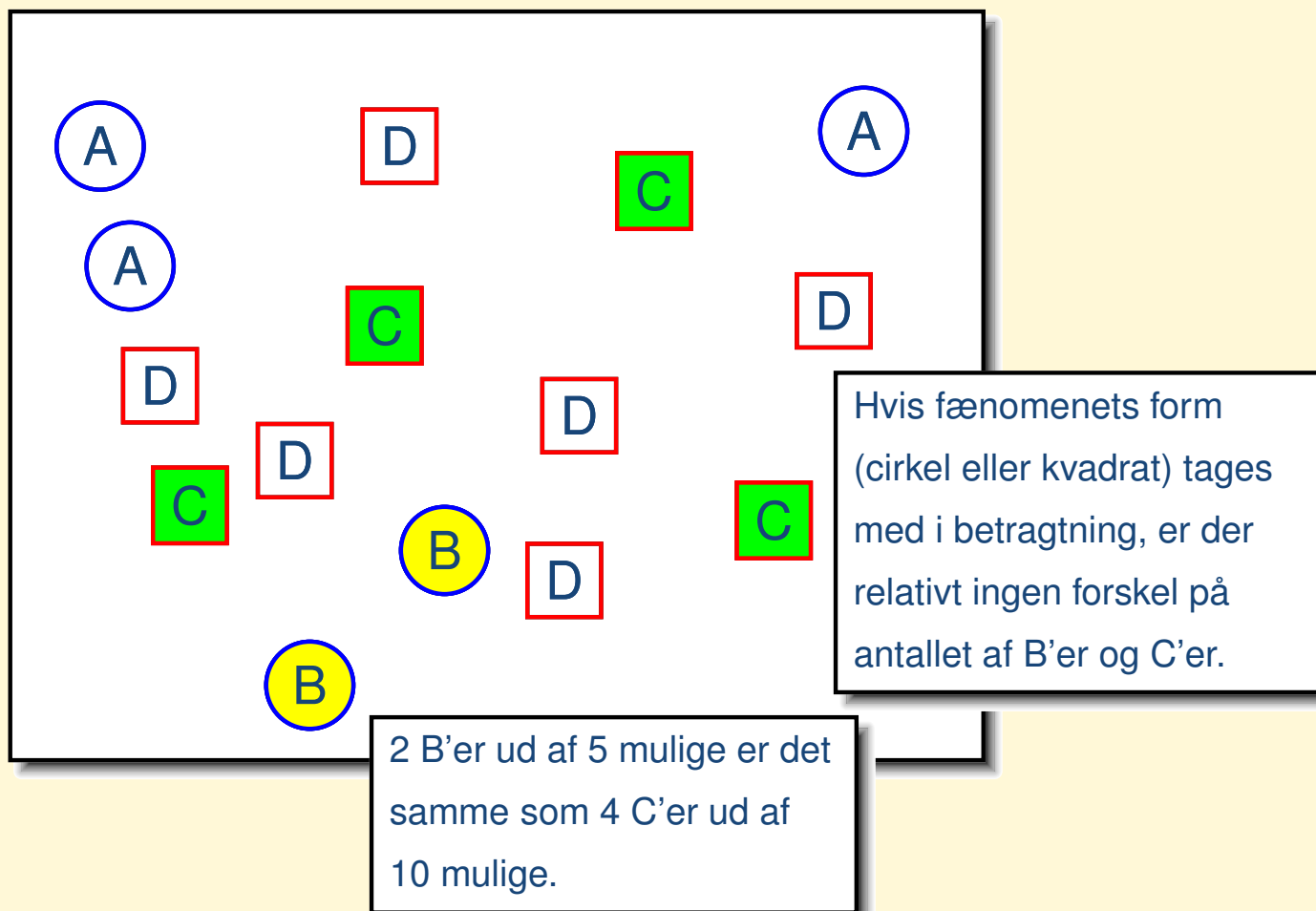
Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Fortolkning af data

Diskuter på baggrund af nedenstående figur udsagnet:
I materialet er der flere forekomster af C end af B.



Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Ordhyppigheder

Formhyppigheder

Fortolkning af data

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Sammenligninger – hvordan?

Opgave 4.1 – 4.3 (30 minutter)

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk
 - ◆ kvantitative udsagn skal være statistisk velbegrundede...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk
 - ◆ kvantitative udsagn skal være statistisk velbegrundede...
 - ◆ og de skal være komplette (ikke kun vedrøre nogle få udvalgte fænomener)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk
 - ◆ kvantitative udsagn skal være statistisk velbegrundede...
 - ◆ og de skal være komplette (ikke kun vedrøre nogle få udvalgte fænomener)
- *vokabularundersøgelser* betyder, at vi her kun interesserer os for isolerede ord, fordi...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk
 - ◆ kvantitative udsagn skal være statistisk velbegrundede...
 - ◆ og de skal være komplette (ikke kun vedrøre nogle få udvalgte fænomener)
- *vokabularundersøgelser* betyder, at vi her kun interesserer os for isolerede ord, fordi...
 - ◆ vi her kun principielt vil demonstrere bestemte kvantitative metoder inden for korpuslingvistikken

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Vokabularundersøgelser

I det følgende skal der gennemføres vokabularundersøgelser på en række tekster. Disse undersøgelser skal være...

- ‘rent’ *korpuslingvistiske*, ikke ‘blot’ *korpusbaserede*, dvs.
 - ◆ det, vi interesserer os for, skal kunne bestemmes algoritmisk
 - ◆ kvantitative udsagn skal være statistisk velbegrundede...
 - ◆ og de skal være komplette (ikke kun vedrøre nogle få udvalgte fænomener)
- *vokabularundersøgelser* betyder, at vi her kun interesserer os for isolerede ord, fordi...
 - ◆ vi her kun principielt vil demonstrere bestemte kvantitative metoder inden for korpuslingvistikken
 - ◆ – men metoderne kan sagtens bruges til at håndtere mere komplekse sproglige størrelser også, forudsat de kan bestemmes algoritmisk (eksempel: syntagmer)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempelmaterialet

Fork.	Udtryk	Tekst	Type if. DDOC	Længde
S1	skrift	Kirsten Fink og Ole Terney: <i>Sådan reguleres genteknologi. Praksis og erfaringer</i> . Foreningen af Bioteknologiske Industrier i Danmark, 1988. Kapitel 1-3.	Fagsproglig monografi om emnet biologi.	5511 ord
S2	skrift	Vita Andersen: <i>Petruskas laks</i> . Gyldendal, 1989. Side 5-8, 12-17 og 22-26.	Roman for børn.	5047 ord
T1	tale	Tekstuddrag fra <i>Projekt Bysociolingvistik</i> fra 1987.	Gruppeinterview med en voksen interviewer og unge interviewede.	5621 ord
T2	tale	Fjernsynsinterview under titlen <i>Gamle mennesker fortæller</i> . DR TV, 1983.	Interviewet handler om historie. Der medvirker en speaker med et par replikker, ellers en interviewer og en interviewet.	5369 ord

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

S1: Sådan reguleres genteknologi

```
<Tekst ID=GxCa>
<p>Kap. 1</p>
{NL}
<p>ANVENDELSE AF GENTEKNOLOGI</p>
{NL}
<p>Hvad er genteknologi?</p>
{NL}
<p>Genteknologi betegner en samling teknikker, der er specielt beregnede til at ændre organismers arveegenskaber på målrettet måde.</p>
<p>Traditionelt forædlingsarbejde har de samme mål, men må bruge metoder, der er mere tidskrævende, og som ofte gør det svært eller umuligt at overskride artsbarriererne.</p>
<p>De nye genteknologi-metoder tillader derimod overførsel af arveegenskaber mellem principielt alle organismer.</p>
<p>Baggrunden for dette er, at alle levende organismer fra bakterier til mennesket indeholder arvemasse, der er opbygget på principielt samme måde. De arvelige egenskaber findes i lange, trådformede molekyler. Det kemiske navn for dette arvemateriale er DNA <note>DNA. På engelsk: Deoxyribonucleic acid, dvs. deoxyribonukleinsyre.</note>.</p>
<p>I begyndelsen af 1970'erne fandt man ud af metoder til at klippe DNA-tråden over på bestemte steder og at klæbe stykkerne sammen igen på en ny måde. Denne rekombination af arvemateriale kan celler foretage naturligt, men det havde ikke tidligere kunnet gøres i laboratoriet.</p>
<p>Rekombineret DNA betegnes også rDNA. Teknikken betegnes undertiden rekombinant-DNA-teknik eller gensplejsning.</p>
<p>Den traditionelle forædlingsteknik benytter sig som sagt også af rekombination af arvemateriale. Når en kornart krydses med en vild, nærtbeslægtet plante, håber man netop på at opnå en gunstig rekombination.</p>
<p>Den nye genteknologi tillader, at man i højere grad specifikt kan styre og kontrollere, hvad der sker. Og den tillader, at et arveanlæg fra en fremmed art indsættes i en organismes arvemasse.</p>
{NL}
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

S2: Petruskas laksko

```
<Tekst ID=RREo>
<p> <kommentar>side 5-8:</kommentar> </p>
{NL}
<p> <f>Hvordan Petruska fik sine laksko</f> </p>
{NL}
<p>Midt om natten vågnede Marie og løb ind til far og mor. Hun krøb op i deres seng og lagde sig på mor.</p>
<p>- Mor, der er en lyd, sagde hun, - det kan være en alf eller en tyveknægt, der vil tage vores ting. Eller måske er det drillenissen, eller spøgelse.</p>
<p>Mor vågnede og lyttede.</p>
<p>- Ja, der er noget der pusler. Nu skal jeg se efter.</p>
<p>- Jeg er bange, sagde Marie.</p>
<p>- Det skal du ikke være, der er ingen der kan komme ind.</p>
<p>- Et spøgelse kan komme ind, sagde Marie, - jeg har selv set det.</p>
<p>- Der findes ingen spøgelser, sagde mor og kyssede hende. - Put dig her, og pas på du ikke vækker far, hviskede hun.</p>
<p>Det, der puslede, var ikke inde i Maries og Petruskas værelse, og heller ikke inde i stuen.</p>
<p>Men ude i køkkenet stod Petruska oppe på spisebordet og smed rosiner op i luften. Hun havde rosiner i håret og på skuldrene og på nattøjet, hendes bare fødder var næsten dækket med rosiner.</p>
<p>- Jeg laver snevejr, sagde Petruska og grinede til mor.</p>
<p>Man skulle skynde sig at grine, når man lavede noget de voksne ikke ville have. Så grinede de sommetider med, og så fik man ikke skældud. Petruska hadede at få skældud. Heldigvis grinede mor.</p>
<p>- Din tossede Petruska, sagde hun, - det er midt om natten.</p>
<p>- Men det er min fødselsdag, sagde Petruska. - Og på min fødselsdag skal der være sne. Jeg kiggede ud ad vinduet og det var så mørkt, men der var ingen sne. Så måtte jeg selv lave det. Snevejr er godt vejr, og når vejret er godt, så har man været sød.</p>
<p>- Tillykke med fødselsdagen, sagde mor, - tillykke med de fem år.</p>
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritmisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

T1: BySoc (DDOC-repræsentation)

```
<Tekst ID=HQUa> <replik id=INT> <p>din broder som bor her hva- hvad hedder han</p> </replik>
<replik id=MD1> <p>Steen</p> </replik>
<replik id=INT> <p>hn{pause} og det er ham der er syvogtyve?</p> </replik>
<replik id=MD1> <p>otteogtyve</p> </replik>
<replik id=INT> <p>otteogtyve {pause}</p> </replik>
<replik id=MD1> <p>jeg har også en anden broder der er femogtyve {pause}</p> </replik>
<replik id=INT> <p>nå</p> </replik>
<replik id=MD1> <p>men han er {tøven} han er en uge indenfor soldaterne {pause} han er lige blevet ind for
soldaterne, ikke?</p> </replik>
<replik id=INT> <p>hjnå</p> </replik>
<replik id=MD1> <p>altså {tøven} {uf} hvad hedder det {latter}</p> </replik>
<replik id=MD2> <p>{uf} genindkaldt</p> </replik>
<replik id=MD1> <p>genindkaldt</p> </replik>
<replik id=INT> <p>njå {pause}</p> </replik>
<replik id=MD1> <p>så til ham {tøven} kan vi ikke {uf} foreløbig så</p> </replik>
<replik id=INT> <p>hvor {tøven} i {tøven} Søværnet?</p> </replik>
<replik id=MD1> <p>usikker>næe det er lige</usikker> ovre på Holmen {pause} i Søværnet {pause} lige der hvor jeg
er {latter} {pause}</p> </replik>
<replik id=INT> <p>jaer. jaer. men plejer I at tænke på Holmen som lige derovre {pause}</p> </replik>
<replik id=MD1> <p>{pause} jaer {pause}</p> </replik>
<replik id=INT> <p>hvor- hvor- hvordan kommer jeres fader på arbejde?</p> </replik>
<replik id=MD1> <p>han {tøven} tager cyklen klokken halv fem om morgenen {pause} så tager han en tur ud over
Knippelsbro og den vej {latter} {pause}</p> </replik>
<replik id=INT> <p>men ka- kan er der ikke en båd man kan tage dernede nu?</p> </replik>
<replik id=MD1> <p>jo men den kører først klokken syv om morgenen</p> </replik>
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

T1: BySoc (partitur-repræsentation)

```
1>                mm ££                S%%%%
2>
A>tten B%%%%%%%%      din bror som bor her hva- hvad hedder han      mm £ og det
K>
-----
1>                otteogtyve                jeg har også en anden bro
2>
A> er ham der er syvogtyve ?                otteogtyve £££
K>
-----
1>r der er femogtyve £      men han er~ ££ han er en uge indenfor soldaterne £ han
2>
A>                nå
K>
-----
1> er lige blevet ind for soldaterne ik'      altså~ £ (fu) hvad hedder det (lat
2>
A>                hjnå
K>
-----
1>ter)                genindkaldt                så til ham~ kan vi ikke (uf) foreløbi
2>      (uf) genindkaldt
A>                nå~ ££
K>
-----
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempelmaterialet

T1: BySoc (partitur-repræsentation)

notation i DDOC

ingen anonymisering

broder

S%%%

hn

{pause}

A>tten B%%%%%%%% din bror som bor her hva- hvad hedder han mm £ og det
K>

broder

1> otteogtyve jeg har også en anden bro

2>

A> er ham der er syvogtyve ? otteogtyve £££

K>

indenfor → inden for?

1>r der er femogtyve £ men han er~ ££ han er en uge indenfor soldaterne £ han

2>

A> nå

K>

ikke

{tøven}

{uf}

1> er lige blevet ind for soldaterne ik' altså~ £ (fu) hvad hedder det (lat

2>

A> hjnå

K>

{uf}

1>ter) genindk njå {pause} ~ kan vi ikke (ur) roreløbi

2> (uf) genindkaldt

A>

nå~ ££

K>

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

T2: Gamle mennesker fortæller

```
<Tekst ID=QydZ>
<replik id=SPE>
<p>I den næste film, Jeg kom fire år for tidligt, kan I høre en ældre dame, Paula Skorup, fortælle om sin barndom i
begyndelsen af århundredeskiftet.</p> </replik>
<kommentar> <p>Klavermusik</p> </kommentar>
<replik id=SKO> <p>Ja, det startede jo sådan, at jeg kom jo fire år for tidligt, og min far og mor de blev ikke gift før
jeg var fire år, da vi skulle have min næste søster.</p> </replik>
<replik id=INT> <p>Ja, det var jo ikke godt den gang.</p> </replik>
<replik id=SKO> <p>Nej, det var det sandelig ikke da. Men min mor boede jo hjemme ved sine forældre, og jeg er jo
født der oppe i møllen oppe i Tindrup. Og brud men der var sandelig, de havde sandelig et barnebarn før også, og
han var ti år den gang jeg var der. Og det var herremanden.</p> </replik>
<replik id=INT> <p>Men, når man øh, skal tænke sådan lidt på Møllegården der, hvordan var det at være der?</p>
</replik>
<replik id=SKO> <p>Ja, se det kan jeg slet ikke huske. For jeg var jo ikke ret gammel, da de flyttede ned til
Tindrup.</p> </replik>
<replik id=INT> <p>Nå.</p> </replik>
<replik id=SKO> <p>Ja, flyttede ned til Tindrup by. Husk på, jeg har jo været ung lidt for længe også.</p> </replik>
<replik id=INT> <p>Ja, det er jo det. Når du nu siger, at du var ung lidt for længe, hvad mener du så egentlig med
det?</p> </replik>
<replik id=SKO> <p>Ja, jeg er jo blevet gammel, ikke. Men du vil jo gerne snakke med gamle koner.</p> </replik>
<replik id=INT> <p>Ja. Herman og dig I var hjemme hos bedstefar og bedstemor.</p> </replik>
<replik id=SKO> <p>Ja, og der havde de sådan et lille hus dernede, men en stor have, men den var delt i to, for det
stod som vejen var gået gennem haven.</p> </replik>
<replik id=INT> <p>Så var der have på den anden side af vejen også.</p> </replik>
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammেলigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelingning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof
 - *kvalitets-produkt, sprog- og litteraturselskab, f.eks., 12.000,-, Alex'*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof
 - *kvalitets-produkt, sprog- og litteraturselskab, f.eks., 12.000,-, Alex'*
- ◆ sær- og samskrevne ord

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof
 - *kvalitets-produkt, sprog- og litteraturselskab, f.eks., 12.000,-, Alex'*
- ◆ sær- og samskrevne ord
 - *kvalitets produkt, f. eks., underskrive vs. skrive under*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof
 - *kvalitets-produkt, sprog- og litteraturselskab, f.eks., 12.000,-, Alex'*
- ◆ sær- og samskrevne ord
 - *kvalitets produkt, f. eks., underskrive vs. skrive under*
- ◆ store/små bogstaver

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Algortimisk ord-definition

Ord: tegnsekvens, der afgrænses af mellemrum i en skreven tekst

■ Problemer:

- ◆ interpunktionstegn m.m., fx punktum, bindestreg, apostrof
 - *kvalitets-produkt, sprog- og litteraturselskab, f.eks., 12.000,-, Alex'*
- ◆ sær- og samskrevne ord
 - *kvalitets produkt, f. eks., underskrive vs. skrive under*
- ◆ store/små bogstaver
 - *man vs. Man, hans vs. Hans*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ **Eksempler:**

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt* og *kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt* og *kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?
- ◆ *Holmen* eller *H%%%%%*?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?
- ◆ *Holmen* eller *H%%%%%*?
- ◆ *introduktion af ordgrænser ved nedskrift af tale* – *hvad er den mest hensigtsmæssige transskription?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?
- ◆ *Holmen* eller *H%%%%%*?
- ◆ *introduktion af ordgrænser ved nedskrift af tale – hvad er den mest hensigtsmæssige transskription?*
 - ord?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?
- ◆ *Holmen* eller *H%%%%%*?
- ◆ *introduktion af ordgrænser ved nedskrift af tale – hvad er den mest hensigtsmæssige transskription?*
 - ord?
 - morfer?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ortografisk normalisering

Hvordan skal man forholde sig til ortografisk normalisering af sit materiale?

■ Eksempler:

- ◆ *kvalitets produkt og kvalitets-produkt* → *kvalitetsprodukt*?
- ◆ *fader* eller *far*?
- ◆ *inden for* eller *indenfor*?
- ◆ *Holmen* eller *H%%%%%*?
- ◆ *introduktion af ordgrænser ved nedskrift af tale – hvad er den mest hensigtsmæssige transskription?*
 - ord?
 - morfer?
 - foner?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelingning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum
 - ◆ flere på hinanden følgende punktummer eller bindestreger

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum
 - ◆ flere på hinanden følgende punktummer eller bindestreger
 - ◆ højre- og venstreparenteser

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum
 - ◆ flere på hinanden følgende punktummer eller bindestreger
 - ◆ højre- og venstreparenteser
 - ◆ dobbelte anførselstegn

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum
 - ◆ flere på hinanden følgende punktummer eller bindestreger
 - ◆ højre- og venstreparenteser
 - ◆ dobbelte anførselstegn
 - ◆ enkelte anførselstegn forudgået eller efterfulgt af mellemrum

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tokeniserens funktion

Tokeniser, som anvendes i de følgende forsøg, opererer med flg. ord-definition:

- ord er de karaktersekvenser, som afgrænses af mellemrum
- følgende tegn fjernes helt:
 - ◆ punktum, udråbstegn, spørgsmålstegn, komma, semikolon, bindestreg efterfulgt af mellemrum
 - ◆ flere på hinanden følgende punktummer eller bindestreger
 - ◆ højre- og venstreparenteser
 - ◆ dobbelte anførselstegn
 - ◆ enkelte anførselstegn forudgået eller efterfulgt af mellemrum
 - ◆ ord omgivet af < og > hhv. { og } (tags)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Tekstrepræsentation, tagget

```
<s s_id="WBNHKKZE" s_nr="471465" txt_id="GxCa" preom="0" bop="1" eop="1">
Traditionelt      [traditionel]      ADJ      NEU S IDF NOM    @>N
forædlingsarbejde [forædlingsarbejde]      N        NEU S IDF NOM    @SUBJ>
har               [have]           <mv>      V        PR AKT             @FMV
de               [den]             ART       nG P DEF           @>N
samme            [samme]             DET       nG nN NOM          @>N
mål              [mål]             N         NEU P IDF NOM      @<ACC
$,
men              [men]             <co-acc>  KC              @CO
må               [måtte]           <aux>     V        PR AKT             @FAUX
bruge            [bruge]           <mv>      V        INF AKT             @ICL-AUX<
metoder          [metode]           N         UTR P IDF NOM      @<ACC
$,
der              [der]             <rel>     INDP      nG nN NOM          @SUBJ>
er               [være]           <mv> <np-close> V        PR AKT             @FS-N<
mere             [meget]           <aquant>  ADV        COM              @>A
tidskrævende     [tidskrævende]     ADJ       nG nN nD NOM      @<SC
$,
og               [og]             <co-fin>  KC              @CO
som              [som]             <rel>     INDP      nG nN              @SUBJ>
ofte             [ofte]           <atemp>   ADV              @ADVL>
gør              [gøre]           <mv> <np-close> V        PR AKT             @FS-N<
det              [den]            PERS      NEU 3S ACC        @F-<ACC
svært            [svær]           ADJ       NEU S IDF NOM      @<OC
eller            [eller]           <co-oc>   KC              @CO
umuligt          [umulig]         ADJ       NEU S IDF NOM      @<OC
at               [at]             INFM              @INFM
overskride       [overskride]     <mv>      V        INF AKT             @ICL-A<
artsbarriererne [art+barriere]   <compound> N         UTR P DEF NOM      @<ACC
$.
</s>
```

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Type- og tokenstatistik

Simple ordoptællinger i eksempelteksterne

Tekst	Tokens	Types	TTF
S1	5511	1822	3,02
S2	5047	1097	4,64
T1	5621	990	5,68
T2	5369	989	5,43

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algotimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types
Statistisk signifikans
Signifikanstest
Resultater
Sammeligning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20 (11)	1,90 (6)	4,80 (1)	3,39 (2)	1,59 (6)
ikke	0,89 (18)	1,88 (7)	3,82 (2)	1,92 (8)	1,00 (15)
der	1,43 (9)	0,69 (26)	3,10 (3)	2,31 (6)	1,14 (12)
jeg	0,02 (1193)	1,49 (11)	2,40 (6)	3,32 (3)	0,52 (26)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20				1,59 (6)
ikke	0,89				1,00 (15)
der	1,43				1,14 (12)
jeg	0,02 (1)				0,52 (26)

Hvordan kan hyppighedsforskellene på *at* og *og* mellem tale- og skriftsprogstekster fortolkes?
 ⇒ Er koordination mere udbredt i talesprog, subordination mere udbredt i skriftsprog?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	Hvad med hyppighedsforskellene for <i>det</i> ? $\Rightarrow$ Mere deiksis i talesprog?			(2)
er	2,27 (3)				(4)
og	1,67 (6)				(1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20 (11)	1,90 (6)	4,80 (1)	3,39 (2)	1,59 (6)
ikke	0,89 (18)	1,88 (7)	3,82 (2)	1,92 (8)	1,00 (15)
der	1,43 (9)	0,69 (26)	3,10 (3)	2,31 (6)	1,14 (12)
jeg	0,02 (1193)	1,49 (11)	2,40 (6)	3,32 (3)	0,52 (26)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2		0 (-)	0,00 (14103)
det	1,20 (11)	1		3,39 (2)	1,59 (6)
ikke	0,89 (18)	1		1,92 (8)	1,00 (15)
der	1,43 (9)	0,69 (26)	5,10 (5)	2,31 (6)	1,14 (12)
jeg	0,02 (1193)	1,49 (11)	2,40 (6)	3,32 (3)	0,52 (26)

Frekvent brug af
jeg i T2 $\Rightarrow$
Person, der ta-
ler om sig selv?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20 (11)	1,90 (6)	4,80		
ikke	0,89 (18)	1,88 (7)	3,82		
der	1,43 (9)	0,69 (26)	3,10		
jeg	0,02 (1193)	1,49 (11)	2,40		

Frekvent brug af
var i T2 $\Rightarrow$ Hand-
ler teksten om no-
get fortidigt?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Hyppigste types

Sammenligning af de tre hyppigste types i hver af eksempelteksterne

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20 (11)	1,90 (6)	4,80 (1)	3,39 (2)	1,59 (6)
ikke	0,89 (18)	1,88			00 (15)
der	1,43 (9)	0,69			14 (12)
jeg	0,02 (1193)	1,49			52 (26)

Frekvent brug af *Petruska* i T2 $\Rightarrow$ Handler teksten om en person af samme navn?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Statistisk signifikans

Endelig kan man med udgangspunkt vores K90-K2000-sammenligninger spørge. . .

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Statistisk signifikans

Endelig kan man med udgangspunkt vores K90-K2000-sammenligninger spørge. . .

- om frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok til, at man kan fortolke dem som indikatorer for sproglig forandring
- om *kambrium* med sikkerhed *ikke* kan fortolkes som indikator for en sproglig forandringsproces

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Statistisk signifikans

Endelig kan man med udgangspunkt vores K90-K2000-sammenligninger spørge. . .

- om frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok til, at man kan fortolke dem som indikatorer for sproglig forandring
- om *kambrium* med sikkerhed *ikke* kan fortolkes som indikator for en sproglig forandringsproces
- om hyppighedsforskelle for lemmaerne *bil*, *land*, *Danmark*, *cykel*, *hus* og *mand*, som kan konstateres mellem K90 og K2000, har konsekvenser for sammenlignende undersøgelser mellem de to korpora

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Statistisk signifikans

Endelig kan man med udgangspunkt vores K90-K2000-sammenligninger spørge. . .

- om frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok til, at man kan fortolke dem som indikatorer for sproglig forandring
- om *kambrium* med sikkerhed *ikke* kan fortolkes som indikator for en sproglig forandringsproces
- om hyppighedsforskelle for lemmaerne *bil*, *land*, *Danmark*, *cykel*, *hus* og *mand*, som kan konstateres mellem K90 og K2000, har konsekvenser for sammenlignende undersøgelser mellem de to korpora
- om der kan opstilles en komplet liste over samtlige lemmaer, der i deres udbredelse i de to korpora afviger så meget fra hinanden, at der næppe kan være tale om et tilfælde.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Statistisk signifikans

Endelig kan man med udgangspunkt vores K90-K2000-sammenligninger spørge. . .

- om frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok til, at man kan fortolke dem som indikatorer for sproglig forandring
- om *kambrium* med sikkerhed *ikke* kan fortolkes som indikator for en sproglig forandringsproces
- om hyppighedsforskelle for lemmaerne *bil*, *land*, *Danmark*, *cykel*, *hus* og *mand*, som kan konstateres mellem K90 og K2000, har konsekvenser for sammenlignende undersøgelser mellem de to korpora
- om der kan opstilles en komplet liste over samtlige lemmaer, der i deres udbredelse i de to korpora afviger så meget fra hinanden, at der næppe kan være tale om et tilfælde.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algotimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Log likelihood (G^2) er en meget brugt test inden for korpuslingvistikken

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammeligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Log likelihood (G^2) er en meget brugt test inden for korpuslingvistikken

χ^2 derimod egner sig mindre godt til korpuslingvistiske formål

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Log likelihood (G^2) er en meget brugt test inden for korpuslingvistikken

χ^2 derimod egner sig mindre godt til korpuslingvistiske formål

Log likelihood er...

- let at bruge
- lavet mhp. sammenligning af vokabular o.l.
- *ikke* helt optimal til sammenligning af store (inhomogene) tekstsamlinger
- *ikke* helt ligetil rent matematisk

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algoritisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans

Signifikanstest

Resultater
Sammelnigning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Log likelihood (G^2) er en meget brugt test inden for korpuslingvistikken

χ^2 derimod egner sig mindre godt til korpuslingvistiske formål

Log likelihood er...

- let at bruge
- lavet mhp. sammenligning af vokabular o.l.
- *ikke* helt optimal til sammenligning af store (inhomogene) tekstsamlinger
- *ikke* helt ligetil rent matematisk

Log likelihood beror som alle statistiske signifikanstest på grundantagelsen:

- at der nok ikke er forskel på de data, man vil sammenligne:
nul-hypotese

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammenligning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og T1 mest signifikant fra hinanden

rang	type	»S1	»T1	G^2
1	jeg		+	174,1
2	så		+	145,4
3	det		+	128,9
4	ikke		+	109,4
5	var		+	100,3
6	sådan		+	82,6
7	jaer		+	81,9
8	af	+		78,3
9	at	+		72,8
10	altså		+	71,9
11	du		+	71,2
12	hun		+	69,9

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikantest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og T1 mest signifikant fra hinanden

rang	type	»S1	»T1	G^2
13	vi		+	65,1
14	for	+		56,7
15	mikroorganismer	+		44,8
16	organismer	+		39,4
17	noget		+	36,0
18	der		+	34,7
19	men		+	34,7
20	gensplejsede	+		34,1
21	her		+	33,6
22	Det	+		32,0
23	naej		+	30,4
24	som	+		29,7

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og T1 mest signifikant fra hinanden

rang	type	» S1	» T1	G^2
25	han		+	29,2
26	om	+		28,7
27	hedder		+	27,8
28	f.eks	+		26,1
29	havde		+	25,8
30	arbejde	+		25,4

OBS! Vi sammenligner to tekster umiddelbart med hinanden
– ikke over et referencekorpus!

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S1	G^2
1	mikroorganismer		+	295,2
2	f.eks		+	229,6
3	organismer		+	227,2
4	1987		+	199,8
5	gensplejsede		+	197,2
6	udsætning		+	177,5
7	miljøstyrelse		+	132,5
8	1		+	110,7
9	EPA		+	98,8
10	Marcker		+	95,4
11	Bacillus		+	85,4
12	klasse		+	83,3

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritimisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammeligning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S1	G^2
13	bakterie		+	77,4
14	K12-kolibakterien		+	71,5
15	genteknologi		+	70,4
16	10		+	70,2
17	2		+	70,2
18	forsøg		+	68,3
19	3		+	68,0
20	Kjeld		+	66,9
21	subtilis		+	66,2
22	1986		+	65,5
23	K12		+	65,5
24	anvendes		+	64,6

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritimisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S1	G^2
25	bakterier		+	62,7
26	organismen		+	60,4
27	retningslinier		+	59,1
28	amerikanske		+	58,5
29	NAS-rapporten		+	57,2
30	sygdomsorganismer		+	55,9

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algotimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest

Resultater

Sammeligning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S2 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S2	G^2
1	Petruska		+	1932,9
2	Marie		+	1022,0
3	mor		+	557,1
4	sagde		+	454,1
5	laksko		+	452,6
6	lakskoene		+	273,6
7	Petruskas		+	173,6
8	Mor		+	168,8
9	du		+	120,7
10	børnehave		+	105,8
11	osse		+	92,7
12	hun		+	89,3

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritimisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S2 og K90 mest signifikant fra hinanden

rang	type	≫ K90	≫ S2	G^2
13	far		+	80,0
14	gaver		+	77,4
15	børnehaven		+	77,0
16	Maries		+	76,1
17	drillenissen		+	57,9
18	dig		+	57,9
19	jeg		+	56,5
20	lyserøde		+	56,3
21	vågnede		+	55,1
22	Bvadr		+	53,7
23	spurgte		+	48,3
24	så		+	46,9

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikantest

Resultater

Sammenligning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller S2 og K90 mest signifikant fra hinanden

rang	type	»K90	»S2	G^2
25	numsen		+	46,3
26	spøgelset		+	45,1
27	sko		+	44,0
28	chokoladefrøer		+	43,4
29	løb		+	42,8
30	at	+		42,5

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algotimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest

Resultater

Sammelnigning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S2	G^2
1	jaer		+	1096,4
2	naej		+	436,2
3	sådan		+	342,7
4	ikke		+	272,2
5	altså		+	256,8
6	det		+	230,8
7	så		+	222,0
8	njå		+	199,3
9	hn		+	197,3
10	jeg		+	194,0
11	Nyboder		+	189,3
12	najåer		+	182,5

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritimisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammeligning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T1 og K90 mest signifikant fra hinanden

rang	type	»K90	»S2	G^2
13	njaer		+	182,5
14	jamen		+	171,3
15	njaåer		+	165,7
16	njåer		+	165,7
17	jae		+	163,6
18	najå		+	148,8
19	der		+	128,5
20	hedder		+	105,9
21	fru		+	104,9
22	derovre		+	101,1
23	nåer		+	98,6
24	vi		+	97,7

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritimisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammeligning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S2	G^2
25	du		+	89,9
26	hnn		+	81,9
27	Suensonsgade		+	81,9
28	Tyttebær-Maja		+	81,9
29	hvad		+	81,7
30	moder		+	80,3

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algoritisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest

Resultater

Sammelnigning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T2 og K90 mest signifikant fra hinanden

rang	type	»K90	»S2	G^2
1	Ja		+	435,2
2	jeg		+	348,5
3	jo		+	341,6
4	så		+	248,7
5	sådan		+	207,3
6	var		+	205,9
7	Rodskov		+	149,7
8	ja		+	132,7
9	da		+	116,4
10	hm		+	112,5
11	han		+	112,1
12	sæbe		+	95,6

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T2 og K90 mest signifikant fra hinanden

rang	type	»K90	»S2	G^2
13	far		+	93,4
14	huske		+	85,1
15	det		+	79,7
16	havde		+	78,6
17	dernede		+	71,2
18	Tindrup		+	65,7
19	bedstemor		+	64,2
20	at	+		61,4
21	du		+	60,5
22	flyttelæssset		+	56,1
23	oppe		+	54,4
24	som		+	51,5

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
 Eksempelmaterialet
 Algoritisk ord-definition
 Ortografisk normalisering
 Tokeniserens funktion
 Tekstrepræsentation, tagget
 Type- og tokenstatistik
 Hyppigste types
 Statistisk signifikans
 Signifikanstest

Resultater

Sammelnigning K90-K2000
 Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Ord, der adskiller T2 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S2	G^2
25	bedstefar		+	50,9
26	henne		+	50,9
27	der		+	49,9
28	nede		+	49,6
29	gangbrættet		+	49,2
30	persillesovs		+	48,0

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algotimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest

Resultater

Sammেলigning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Sammelnigning K90-K2000

Log-likelihood-sammenligning af K2000 og K90

rang	type	≫ K2000	≫ K90	G^2
1	EU	+		11126
2	du		+	10887
3	\$	+		10076
4	var		+	7849
5	hun		+	6731
6	mio.	+		6612
7	at	+		6379
8	havde		+	6004
9	\$1999	+		5788
10	\$1998	+		5698
11	ham		+	5299
12	siger	+		5161

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algortimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Sammelnigning K90-K2000

Log-likelihood-sammenligning af K2000 og K90

rang	type	≫ K2000	≫ K90	G^2
13	\$1997	+		4909
14	EF		+	4394
15	hende		+	4227
16	%	+	+	4146
17	han		+	3251
18	jeg		+	3243
19	NN		+	3075
20	mill.		+	3027
21	Du		+	2938
22	\$1996	+		2877
23	dig		+	2876
24	danske	+		2847

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algotimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest
Resultater

Sammelnigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Sammেলigning K90-K2000

Log-likelihood-sammenligning af K2000 og K90

rang	type	≫ K2000	≫ K90	G^2
25	\$2000	+		2821
26	pct.	+		2789
27	direktør	+		2742
28	Hun		+	2689
29	har	+		2663
30	Internettet	+		2642
31	mig		+	2601
32	EUs	+		2514
33	\$2001	+		2481
34	Dansk=Folkeparti	+		2474
35	Ytring	-	+	2448

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser

Eksempelmaterialet

Algoritimisk ord-definition

Ortografisk normalisering

Tokeniserens funktion

Tekstrepræsentation, tagget

Type- og tokenstatistik

Hyppigste types

Statistisk signifikans

Signifikanstest

Resultater

Sammেলigning K90-K2000

Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

G^2 -værdier for udvalgte ord i K2000 over for K90

lemma	K2000	K90	G^2	overrepræsenteret ($p \geq 0,99$)
mobiltelefon	1486	59	1607	i K2000
benchmarking	34	0	40	i K2000
biltelefon	9	51	33	i K90
kambrium	0	4	4	nej
bil	8353	10364	265	i K90
land	28204	21455	769	i K2000
Danmark	30677	22217	1168	i K2000
cykel	1343	1773	69	i K90
hus	8146	12016	840	i K90
mand	24612	29878	639	i K90

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Vokabularundersøgelser
Eksempelmaterialet
Algortimisk ord-definition
Ortografisk normalisering
Tokeniserens funktion
Tekstrepræsentation, tagget
Type- og tokenstatistik
Hyppigste types
Statistisk signifikans
Signifikanstest
Resultater
Sammelnigning K90-K2000
Konklusioner

Struktur og kollokation

Heuristisk tekstklassifikation

Litteratur m.m.

Struktur og kollokation

Opgave 5.1 (10 minutter)

Frekvensprofiler og struktur

- Frekvensprofiler har vi
 - ◆ hidtil kun anvendt på enkeltord
 - ◆ men metoden lader sig også anvende på andre sproglige enheder, fx
 - grupper af (på hinanden følgende) ord: *n-grammer*
 - (grupper af) ordklasse-tags
 - (grupper af) syntaktiske tags
 - (grupper af) bogstaver
 - (grupper af) stavelser
 - (grupper af) foner
 - (grupper af) morfer

Ved at opstille frekvensprofiler for forskellige *n*-grammer i et sprog vil man kunne få et indtryk af de kombinatoriske muligheder i sproget

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Frekvensprofilers begrænsing

Men hvilken ulempe har frekvensprofiler over n-grammer?

i præp.	og konj.
og konj.	i præp.
af præp.	til præp.
på præp.	af præp.
til præp.	på præp.
fra præp.	at konj.
at konj.	end præp.
vækst sb.	for præp.
der pron.	nok adv.
med præp.	som præp.
stigning sb.	der pron.
er vb.	med præp.
som pron.	er vb.
for præp.	som pron.
kritik sb.	fra præp.

Hvilken kolonne indeholder ord, der typisk står foran *stærk* hhv. *kraftig*?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsing

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Frekvensprofilers begrænsing

Men hvilken ulempe har frekvensprofiler over n-grammer?

i præp.	og konj.
og konj.	i præp.
af præp.	til præp.
på præp.	af præp.
til præp.	på præp.
fra præp.	at konj.
at konj.	end præp.
vækst sb.	for præp.
der pron.	nok adv.
med præp.	som præp.
stigning sb.	der pron.
er vb.	med præp.
som pron.	er vb.
for præp.	som pron.
kritik sb.	fra præp.

Hvilken kolonne indeholder ord, der typisk står foran *stærk* hhv. *kraftig*?

Løsning:

kraftig til venstre, *stærk* til højre. . .

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsing

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Mutual Information (MI) vil her blive eksemplificeret i forbindelse med fremfinding af iøjnefaldende samforekomster af ord $\implies$ kandidater for faste udtryk m.m.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Mutual Information (MI) vil her blive eksemplificeret i forbindelse med fremfinding af iøjnefaldende samforekomster af ord $\implies$ kandidater for faste udtryk m.m.

Baggrunden for MI

- informationsteori
 - ◆ teori til matematisk beskrivelse af informationshåndtering i tekniske systemer, fx inden for telekommunikation og it
- en hypotetisk forventning om, at alle ord eller andre relevante sproglige enheder i et korpus kommer i vilkårlig, tilfældig rækkefølge:
 - ◆ der er ingen tilbagevendende mønstre eller regelmæssigheder

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Mutual Information (MI) vil her blive eksemplificeret i forbindelse med fremfinding af iøjnefaldende samforekomster af ord $\implies$ kandidater for faste udtryk m.m.

Baggrunden for MI

- informationsteori
 - ◆ teori til matematisk beskrivelse af informationshåndtering i tekniske systemer, fx inden for telekommunikation og it
- en hypotetisk forventning om, at alle ord eller andre relevante sproglige enheder i et korpus kommer i vilkårlig, tilfældig rækkefølge:
 - ◆ der er ingen tilbagevendende mønstre eller regelmæssigheder

Hvordan hænger denne forventning sammen med nul-hypotesen?

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempler: kraftig/stærk

Venstre kollokater for *kraftig* og *stærk* i Korpus 2000 fundet vha. MI

<i>kraftig</i>	<i>stærk</i>
<i>regnskyl</i> sb.	<i>kritisabelt</i> adj.
<i>magnetfelt</i> sb.	<i>psyke</i> sb.
<i>jordskælv</i> sb.	<i>overdrevet</i> sb.
<i>eksplosion</i> sb.	<i>smerter</i> sb.
<i>blæst</i> sb.	<i>syre</i> sb.
<i>vind</i> sb.	<i>kontrast</i> sb.
<i>stigning</i> sb.	<i>vendinger</i> sb.
<i>ryk</i> sb.	<i>sammenhold</i> sb.
<i>reduktion</i> sb.	<i>back</i> sb.
<i>udbygning</i> sb.	<i>kræfter</i> sb.
<i>vækst</i> sb.	<i>følelser</i> sb.
<i>orgasme</i> sb.	<i>personlighed</i> sb.
<i>stød</i> sb.	<i>position</i> sb.
<i>støj</i> sb.	<i>modstander</i> sb.
<i>vendinger</i> sb.	<i>farver</i> sb.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Eksempel fra øvelsen

Skemaet viser MI-beregnedede kollokater for *bevægelse*:

Korpus 2000		Korpus 90	
venstre	højre	venstre	højre
●●● glidende vb.	●●● hånden sb.	●●● nyreligiøse adj.	●●● hurtige adj.
●●● roterende adj.		●●● fejende adj.	
●●● langsomme adj.		●●● sindene sb.	
●●● blide adj.		●●● roterende adj.	
●●● nazistiske adj.		●●● langsomme adj.	
●●● folkelige adj.		●●● rytmiske adj.	
●●● pludselige adj.		●●● folkelige adj.	
●●● rytmiske adj.		●●● revolutionære adj.	
●●● hurtige adj.		●●● glidende vb.	
●●● folkelig adj.		●●● rolige adj.	

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Frekvensprofiler og struktur

Frekvensprofilers begrænsning

Mutual Information

Eksempler: kraftig/stærk

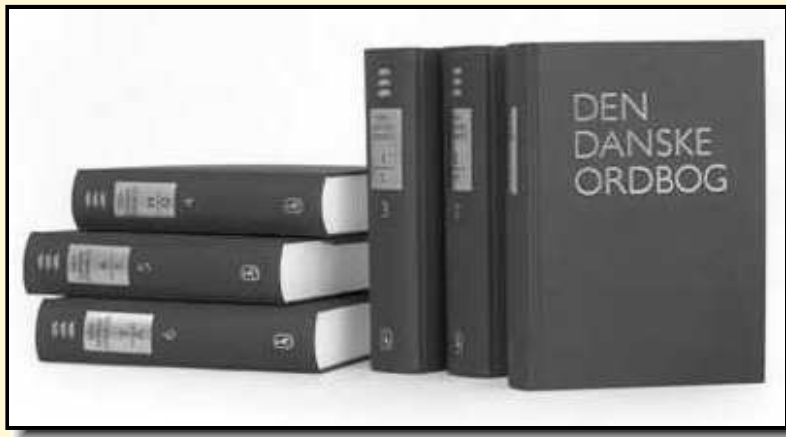
Eksempel fra øvelsen

Heuristisk tekstklassifikation

Litteratur m.m.

Heuristisk tekstklassifikation

Opgave 6.1 (10 minutter)



- Korpusbaseret ordbog over moderne dansk
- Udgivet af DSL 2003-2005 i seks bind
- Integreres i *Ordnet.dk*...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

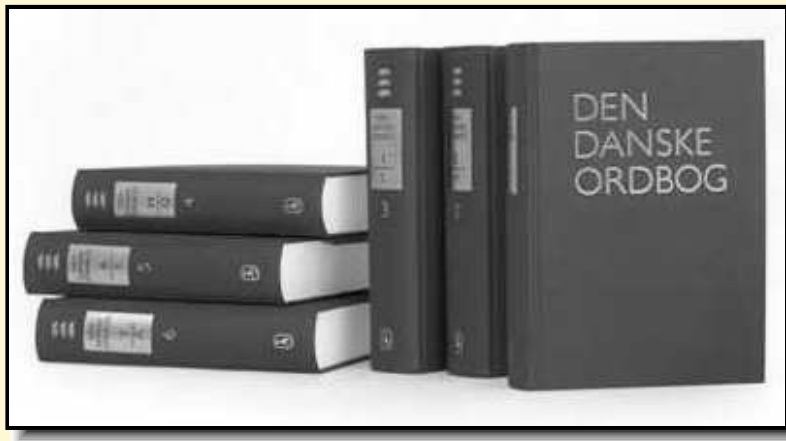
Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.



- Korpusbaseret ordbog over moderne dansk
- Udgivet af DSL 2003-2005 i seks bind
- Integreres i *Ordnet.dk*...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

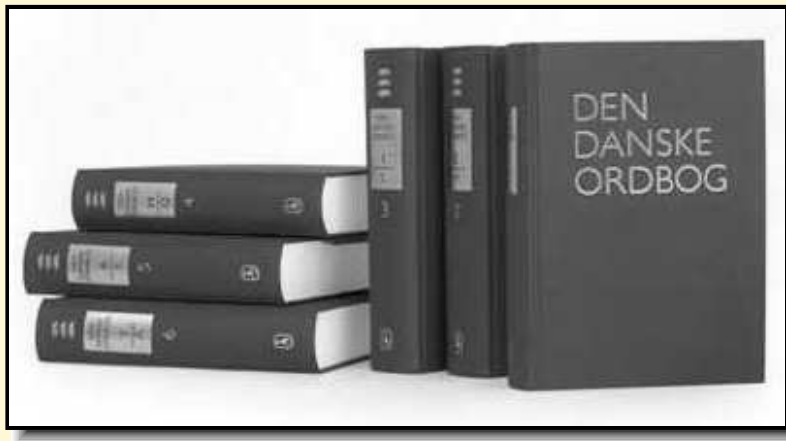
Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.



- Korpusbaseret ordbog over moderne dansk
- Udgivet af DSL 2003-2005 i seks bind
- Integreres i *Ordnet.dk*...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

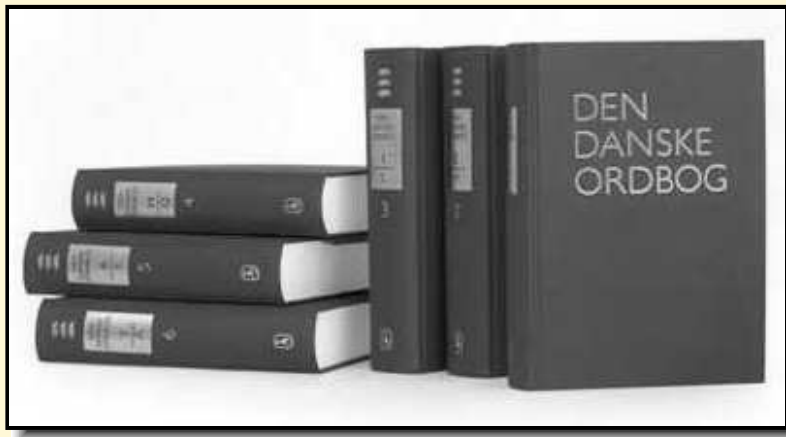
Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.



- Korpusbaseret ordbog over moderne dansk
- Udgivet af DSL 2003-2005 i seks bind
- Integreres i *Ordnet.dk*...
 - ◆ *Ordnet.dk* udvides løbende med nye ord og nye betydninger...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

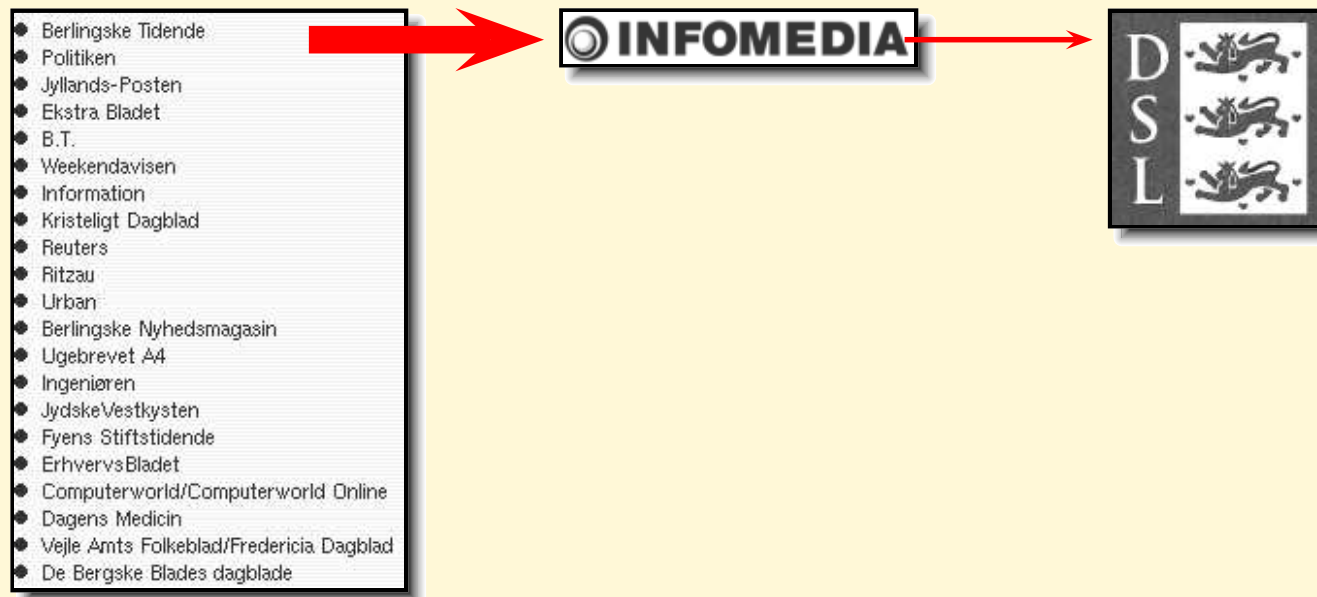
Testning

Konklusion

Litteratur m.m.

Kilder til nye ord

- Til det her beskrevne formål benyttes især avismateriale, leveret gennem *www.infomedia.dk*



Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Før nye ord beskrives i *Ordnet.dk*, deles de op i emnegrupper (domæner):

1. Hver tekst tilordnes et domæne
2. Påfaldende nye ord inden for hvert domæne udtrækkes
3. De udtrukne ord er kandidater til optagelse i *DDO/Ordnet*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner?*
 - (b) indhold – *intensionel definition af et domæne?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner*?
 - (b) indhold – *intensionel definition af et domæne*?

Decimalklassifikationssystemet *DK5*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner?*
 - (b) indhold – *intensionel definition af et domæne?*

Decimalklassifikationssystemet *DK5*

2. Klassifikationsprocedure – *hvordan tilordnes en tekst et bestemt domæne?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner?*
 - (b) indhold – *intensionel definition af et domæne?*

Decimalklassifikationssystemet *DK5*

2. Klassifikationsprocedure – *hvordan tilordnes en tekst et bestemt domæne?*

Heuristisk, baseret på DDO's korpus

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner?*
 - (b) indhold – *intensionel definition af et domæne?*

Decimalklassifikationssystemet *DK5*

2. Klassifikationsprocedure – *hvordan tilordnes en tekst et bestemt domæne?*

Heuristisk, baseret på DDO's korpus

88,6% af tekststykkerne er tilordnet ét af 66 domæner

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

1. En brugbar domæneklassifikation:
 - (a) granularitet – *antal domæner?*
 - (b) indhold – *intensionel definition af et domæne?*

Decimalklassifikationssystemet *DK5*

2. Klassifikationsprocedure – *hvordan tilordnes en tekst et bestemt domæne?*

Heuristisk, baseret på DDO's korpus

88,6% af tekststykkerne er tilordnet ét af 66 domæner

Herfra udtrækkes 66 domænespecifikke vokabularer, som grundlag for klassifikation af ukendt tekst

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Domænespecifikke vokabularer

1. **Opbygning af domænespecifikke subkorpora:**
 - DDOC-domænekoder $\implies$ 66 domænespecifikke subkorpora

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Domænespecifikke vokabularer

1. **Opbygning af domænespecifikke subkorpora:**
 - DDOC-domænekoder $\Rightarrow$ 66 domænespecifikke subkorpora
2. **Opstilling af frekvensprofiler:**
 - DDOC + subkopora $\Rightarrow$ frekvensprofiler

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Domænespecifikke vokabularer

1. **Opbygning af domænespecifikke subkorpora:**
 - DDOC-domænekoder $\implies$ 66 domænespecifikke subkorpora
2. **Opstilling af frekvensprofiler:**
 - DDOC + subkopora $\implies$ frekvensprofiler
3. **Sammenligning af frekvensprofiler:**
 - Hver af de 66 domænespecifikke frekvensprofiler sammenlignes med frekvensprofilen for hele korpusset
Signifikanstest: log likelihood
 - Overrepræsenterede typer ($p \geq 0.99$) inden for et domæne udgør dets domænespecifikke vokabular D .

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Eksempel

D_{edb}	$D_{filosofi}$	$D_{oekonomi}$
data	mennesket	kr
programmer	kierkegaard	X,X
computer	moral	pct
computeren	løgstrup	procent
edb	aristoteles	kroner
computere	filosofi	rente
ibm	fornuft	offentlige
pc	platon	økonomiske
kan	kierkegaards	bank
mb	tim	X
apple	den	økonomi
amiga	menneskets	vil
commodore	filosof	mia

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Metodiske problemstillinger

- Det arbitrære signifikansniveau ($p \geq 0.99$) $\implies$ antal typer i de domænespecifikke vokabularer

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Metodiske problemstillinger

- Det arbitrære signifikansniveau ($p \geq 0.99$) $\implies$ antal typer i de domænespecifikke vokabularer
- De domænespecifikke subkorporer har forskellig størrelse $\implies$ afledte vokabularer er forskelligt store, fx
Folklore: 1957 typer
Sport: 16022 typer
Gennemsnit for alle domæner: 7256 typer

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Metodiske problemstillinger

- Det arbitrære signifikansniveau ($p \geq 0.99$) $\implies$ antal typer i de domænespecifikke vokabularer
 - De domænespecifikke subkorporer har forskellig størrelse $\implies$ afledte vokabularer er forskelligt store, fx
Folklore: 1957 typer
Sport: 16022 typer
Gennemsnit for alle domæner: 7256 typer
- Ved domænetilordningen må der tages højde for de domænespecifikke vokabularers forskellige størrelser

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Metodiske problemstillinger

- Det arbitrære signifikansniveau ($p \geq 0.99$) $\implies$ antal typer i de domænespecifikke vokabularer
- De domænespecifikke subkorporer har forskellig størrelse $\implies$ afledte vokabularer er forskelligt store, fx
Folklore: 1957 typer
Sport: 16022 typer
Gennemsnit for alle domæner: 7256 typer
 $\longrightarrow$ Ved domænetilordningen må der tages højde for de domænespecifikke vokabularers forskellige størrelser
- Højfrekvente funktionsord optræder som signifikante, fx
 - ◆ *den*
 - ◆ *vil*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Metodiske problemstillinger

- Det arbitrære signifikansniveau ($p \geq 0.99$) $\implies$ antal typer i de domænespecifikke vokabularer
- De domænespecifikke subkorporer har forskellig størrelse $\implies$ afledte vokabularer er forskelligt store, fx
Folklore: 1957 typer
Sport: 16022 typer
Gennemsnit for alle domæner: 7256 typer
 $\longrightarrow$ Ved domænetilordningen må der tages højde for de domænespecifikke vokabularers forskellige størrelser
- Højfrekvente funktionsord optræder som signifikante, fx
 - ◆ *den*
 - ◆ *vil* $\longrightarrow$ Fjernes ikke fra de domænespecifikke vokabularer (faste udtryk, kollokationer)

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én
Løsning: største fællesmængde mellem D og mængden af *teksttokens* W

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én
Løsning: største fællesmængde mellem D og mængden af *teksttokens* W
- Domæner med store vokabularer vil have lettere ved at score højt

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én
Løsning: største fællesmængde mellem D og mængden af *teksttokens* W
- Domæner med store vokabularer vil have lettere ved at score højt
Løsning: der må tages højde for størrelsen af de enkelte domænespecifikke vokabularer

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én
Løsning: største fællesmængde mellem D og mængden af *teksttokens* W
- Domæner med store vokabularer vil have lettere ved at score højt
Løsning: der må tages højde for størrelsen af de enkelte domænespecifikke vokabularer
- Funktionsord kan få for høj vægt

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Udgangspunkt:

- Største fællesmængde $|D \cap T|$ mellem et domænespecifikt vokabular D og vokabularet T i den tekst, som skal klassificeres

Overvejelser:

- Højfrekvente domænespecifikke typer tæller kun som én
Løsning: største fællesmængde mellem D og mængden af *teksttokens* W
- Domæner med store vokabularer vil have lettere ved at score højt
Løsning: der må tages højde for størrelsen af de enkelte domænespecifikke vokabularer
- Funktionsord kan få for høj vægt
Løsning: der må tages højde for antallet af domæner et givet token er medlem af

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

2

Denne værdi w skal være omvendt proportional til antallet af domæner d , i hvis specifikke vokabularer tekst-tokenet optræder:

▷ *mange domæner* $\implies$ *lav værdi (og omvendt)*

$$w = \frac{1}{d} \text{ hvor}$$

$$d = \sum_i |t \cap D_i|$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

2

Denne værdi w skal være omvendt proportional til antallet af domæner d , i hvis specifikke vokabularer tekst-tokenet optræder:

▷ *mange domæner* $\implies$ *lav værdi (og omvendt)*

$$w = \frac{1}{d} \text{ hvor } d = \sum_i |t \cap D_i|$$

3

Beregn en vægt v for den samlede score, som er omvendt proportional til 'omtrent' størrelsen af domænets specifikke vokabular D :

▷ *stort vokabular* $\implies$ *lidt vægt (og omvendt)*

$$v = \frac{1}{\sqrt{|D|}}$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

2

Denne værdi w skal være omvendt proportional til antallet af domæner d , i hvis specifikke vokabularer tekst-tokenet optræder:

▷ *mange domæner $\implies$ lav værdi (og omvendt)*

$$w = \frac{1}{d} \text{ hvor } d = \sum_i |t \cap D_i|$$

3

Beregn en vægt v for den samlede score, som er omvendt proportional til 'omtrent' størrelsen af domænets specifikke vokabular D :

▷ *stort vokabular $\implies$ lidt vægt (og omvendt)*

$$v = \frac{1}{\sqrt{|D|}}$$

4

Beregn en vægt for den samlede score, som tager højde for forholdet mellem de **tekst-tokens, som optræder i et af de domænespecifikke vokabularer k** , og dem, som ikke gør u :

▷ *mange 'kendte' tokens $\implies$ høj vægt (og omvendt)*

$$\frac{k}{u}$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

2

Denne værdi w skal være omvendt proportional til antallet af domæner d , i hvis specifikke vokabularer tekst-tokenet optræder:

▷ *mange domæner* $\implies$ *lav værdi (og omvendt)*

$$w = \frac{1}{d} \text{ hvor } d = \sum_i |t \cap D_i|$$

3

Beregn en vægt v for den samlede score, som er omvendt proportional til 'omtrent' størrelsen af domænets specifikke vokabular D :

▷ *stort vokabular* $\implies$ *lidt vægt (og omvendt)*

$$v = \frac{1}{\sqrt{|D|}}$$

4

Beregn en vægt for den samlede score, som tager højde for forholdet mellem de tekst-tokens, som optræder i et af de domænespecifikke vokabularer k , og dem, som ikke gør u :

▷ *mange 'kendte' tokens* $\implies$ *høj vægt (og omvendt)*

$$\frac{k}{u}$$

5

Gør endelig scoren **relativ til** tekstlængden i **antal tokens n**

$$\frac{1}{n}$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Beregning af domæne-score

1

Lad hvert token t fra teksten W , som også er i domænets specifikke vokabular D , addere en bestemt værdi w til scoren

$$\sum_{t \in D \cap W} w_t$$

2

Denne værdi w skal være omvendt proportional til antallet af domæner d , i hvis specifikke vokabularer tokenet t optræder

$$w = \frac{1}{d} \text{ hvor } d = \sum_i |t \cap D_i|$$

3

Be
on
do
▷ s

$$s_D = \frac{1}{n} \cdot \frac{k}{u} \cdot v \cdot \sum_{t \in D \cap W} w_t$$

$$v = \frac{1}{\sqrt{|D|}}$$

4

Beregn en vægt for den samlede score, som tager højde for forholdet mellem de tekst-tokens, som optræder i et af de domænespecifikke vokabularer k , og dem, som ikke gør u :
▷ mange 'kendte' tokens $\Rightarrow$ høj vægt (og omvendt)

$$\frac{k}{u}$$

5

Gør endelig scoren relativ til tekstlængden i antal tokens n

$$\frac{1}{n}$$

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Bestemmelse af nye ord

1. Sammenlign frekvensprofiler vha. log likelihood:
 - nyt domænespecifikt materiale $\longleftrightarrow$ DDO's Korpus
2. fremtrædende ord i det nye materiale $\implies$ kandidater til optagelse i OrdNet

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Eksempel: tekst

Du skal bruge en diskette til installationen. På et tidspunkt bliver du spurgt om du vil lave en bootdiskette. Erfaringen siger at det godt kan betale sig at formatere en diskette i forvejen med tjek for dårlige sektorer. Før du installerer Linux, skal der være en partition til rådighed, der er stor nok til at rumme det hele (samt en swap-partition). I løbet af Linux-installationen vil der blive lejlighed til at repartitionere så meget, du har behov for, inden for den plads, der nu er blevet til rådighed.

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Klassifikation

- Teksten klassificeres som edb-tekst

Sammenligning

- **OBS!** Tekstens lille størrelse vil forvrænge resultatet!
- Men vi laver en sammenligning alligevel . . .
 - ◆ og dets mest fremtrædende ord listes som kandidater
 - ◆ sammen med domæne-/fagtilordninger i DDO's definitioner

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Nye ord?

Type	f_{DDOC}	f_{eks}	DDO-fagkoder
diskette	78	2	<i>edb</i>
bootdiskette	0	1	artikel mangler
formatere	0	1	<i>edb</i>
linux	0	1	artikel mangler
linux-installationen	0	1	artikel mangler
★partition	0	1	artikel mangler
repartitionere	0	1	artikel mangler
swap-partition	0	1	artikel mangler

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Nye betydninger?

Type	f_{DDOC}	f_{eks}	DDO domains
rådighed	1730	2	<i>alment</i>
★installerer	16	1	<i>alment</i> <i>teknik</i>
du	143798	5	<i>alment</i>
★installationen	34	1	<i>teknik</i> <i>kunst</i> <i>militær</i>
tjek	100	1	<i>alment</i>
★sektorer	112	1	<i>samfund</i> <i>politik</i> <i>matematik</i>

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Opgave:

Bestem ny domænespecifikke ord til leksikografisk beskrivelse

Fremgangsmåde:

1. Korpus $\Rightarrow$ domænespecifikke vokabularer
2. Domænespecifikke vokabularer $\Rightarrow$ tekstklassifikation
3. Domæneklassificeret nyt materiale $\longleftrightarrow$ oprindeligt korpus
4. Fremtrædende ord $\Rightarrow$ kandidater for nye artikler/definitioner

Hvert af disse skridt involverer nogle grundlæggende beslutninger som utilsigtet kan have indflydelse på resultatet

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

- **DDO-Korpussets domæneklassifikation**
 - ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

■ DDO-Korpussets domæneklassifikation

- ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne

→ *Færre domæner, mindre mængdeforskel?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

- **DDO-Korpussets domæneklassifikation**
 - ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne
 - *Færre domæner, mindre mængdeforskel?*
- **Signifikanstest**
 - ◆ Log likelihood

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

- **DDO-Korpussets domæneklassifikation**
 - ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne
 - *Færre domæner, mindre mængdeforskel?*
- **Signifikanstest**
 - ◆ Log likelihood
 - *Bedre egnede tests (fx Mann-Whitney ranks-testen)?*
 - *Afspejler de det undersøgte fænomens natur?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

- **DDO-Korpussets domæneklassifikation**
 - ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne
 - *Færre domæner, mindre mængdeforskel?*
- **Signifikanstest**
 - ◆ Log likelihood
 - *Bedre egnede tests (fx Mann-Whitney ranks-testen)?*
 - *Afspejler de det undersøgte fænomens natur?*
- **Klassifikationsproceduren**
 - ◆ Den bør afspejle egenskaber ved teksten
 - Token-overlapning, vokabularstørrelse, 'unikhed', kendte vs. ukendte tokens

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Grundlæggende beslutninger

- **DDO-Korpussets domæneklassifikation**
 - ◆ Stort antal domæner og store forskel på mængden af tekstmateriale for hvert domæne
 - *Færre domæner, mindre mængdeforskel?*
- **Signifikanstest**
 - ◆ Log likelihood
 - *Bedre egnede tests (fx Mann-Whitney ranks-testen)?*
 - *Afspejler de det undersøgte fænomens natur?*
- **Klassifikationsproceduren**
 - ◆ Den bør afspejle egenskaber ved teksten
 - Token-overlapning, vokabularstørrelse, 'unikhed', kendte vs. ukendte tokens
 - *Andre egenskaber, fx G^2 -rangen?*
 - *Intuitionsbaserede egenskaber: konsevenser?*
 - *Er kvantificeringen passende?*
 - ◆ **Ja:** acceptable resultater
 - ◆ **Nej:** hverken forklarer eller afspejler sprogets natur

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

- Gensidige afhængigheder mellem disse beslutninger er komplekse

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

- Gensidige afhængigheder mellem disse beslutninger er komplekse
→ Test af forskellige alternerende parametre

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

- Gensidige afhængigheder mellem disse beslutninger er komplekse
→ Test af forskellige alternerende parametre
- Sådan kan klassifikationsproceduren testes:

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

- Gensidige afhængigheder mellem disse beslutninger er komplekse
→ Test af forskellige alternerende parametre
- Sådan kan klassifikationsproceduren testes:
 - ◆ Del DDOC op i to dele
 - ◆ Samme relative andel tekst fra hvert domæne
 - ◆ Del 1 $\Rightarrow$ domænespecifikke vokabularer
 - ◆ Del 2 $\Rightarrow$ testning

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Godt:

Skidt:

Og...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Godt:

- Metoden er brugbar til opgaven

Skidt:

Og...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Godt:

- Metoden er brugbar til opgaven

Skidt:

- Ingen forklaring på, hvad der gør et ord eller en tekst *domæne-specifik*
- Ingen forklaringen på, hvad der gør et ord til et *nyt ord*

Og...

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Godt:

- Metoden er brugbar til opgaven

Skidt:

- Ingen forklaring på, hvad der gør et ord eller en tekst *domæne-specifik*
- Ingen forklaringen på, hvad der gør et ord til et *nyt ord*

Og...

- Selvom metoden er kvantitativ...
 - ◆ så er den stadig baseret på menneskelige intuition om sproget
 - ◆ ... *Men mon ikke den deler denne skæbne med de fleste kvantitative metoder inden for sprogvidenskaben?*

Program

Grundlæggende

Korpuslingvistik

Finde, tælle, konkludere

Sammenligninger

Struktur og kollokation

Heuristisk tekstklassifikation

DDO og Ordnet.dk

Kilder til nye ord

Opdateringsprocessen

Forudsætninger

Domænespecifikke vokabularer

Eksempel

Metodiske problemstillinger

Tekstklassifikationen

Beregning af domæne-score

Bestemmelse af nye ord

Eksempel: tekst

Eksempel: procedure

Nye ord?

Nye betydninger?

Diskussion af metoden

Grundlæggende beslutninger

Testning

Konklusion

Litteratur m.m.

Litteratur m.m.
Jf. særskilt udleveret liste...

Slut for i dag!
Tak!