Luonnollisen kielen tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2003

Luennot: **Timo Honkela**
Laskuharjoitukset: **Vesa Siivola**

Luentokalvot: Krista Lagus ja Timo Honkela

4. Yleisen kielitieteen perustietoja

4.1 Kielellisen analyysin eri tasoista

Käsiteltäviä kielellisiä yksiköitä

foneemi, morfeemi, sananmuoto, lekseemi, käsite, lause, virke, kappale, dokumentti, korpus

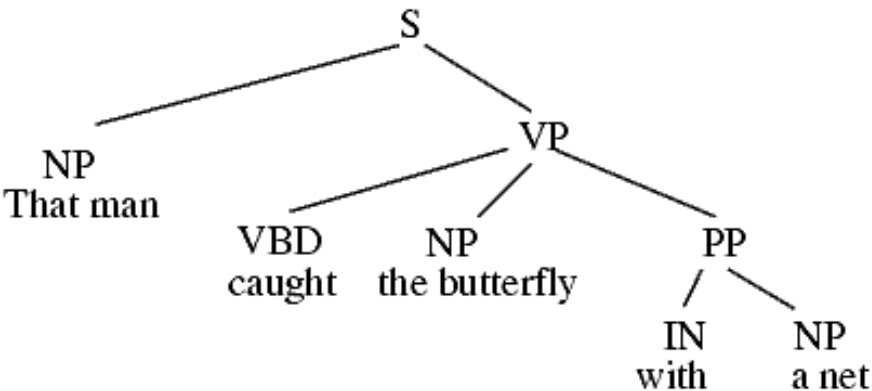
Tiedon lajeja, eri yksiköiden tasoilla

foneettinen ja fonologinen, morfologinen, syntaktinen, semanttinen, pragmaattinen, diskurssitieto, maailmantieto

Esimerkki syntaktisesta analyysista

Käsitteitä: Sanakategoriat, Lauserakennekielioppi, Dependenssikielioppi

Lauserakennekieliopin jäsennyspuu



Esimerkki morfosyntaktisesta analyysistä

Tuotettu Connexorin FDG:llä
FDG=Functional Dependency Grammar
http://www.connexor.com

1	Haetaan	hakea	main: 0 &+MV V PASS IND PRES
2	rahtisatamasta	rahti#satama	sou: 1 &NH N SG ELA
3	lakkaa	lakka	obj: 1 &NH N SG PTV
4	,	,	
5	kun	kun	pm: 6 &CS CS
6	lakkaa	lakata	tmp: 1 &+MV V ACT IND PRES SG3
7	satamasta	sataa	obj: 6 &-MV V ACT INF3 SG ELA
8	.	.	

5. Korpustyöskentely

5.1 Välineitä ja tekniikoita

- Ohjelmointikieliä: Perl, Awk, Python, Prolog
- Tekstihahmojen tunnistus: säännölliset lausekkeet (regexps)
- Sanojen koodaustapoja: sanojen korvaus numeroilla (taulukointi, hajautus (hash table))
- Frekvenssitietojen keräys

5.2 Esimerkki Perl-kielellä

5.3 Kontekstivapaat kieliopit ja Prolog-kieli

- Prolog (PROgrammation LOGique) on nimensä mukaisesti logiikkaohjelmointikieli
- Prolog-kieltä hyödynnetään usein lauseenjäsennyksessä tai perinteisessä tietämyksen kuvaamisessa
- Prolog-kielen laajennuksena on monessa järjestelmässä mukana kontekstivapaiden ja niin sanottujen DFG (Definite Clause Grammar) -kielioppien kehittelyyn sopiva formalismi

5.4 Definite Clause Grammar -esimerkki

```
% DCG-kielioppi

sentence --> noun_phrase, verb_phrase.

noun_phrase --> determiner, noun.
noun_phrase --> noun.

verb_phrase --> verb.
verb_phrase --> verb, noun_phrase.
```

5.5 Definite Clause Grammar -esimerkki, jatkoa

```
% Sanasto

determiner --> [the].
determiner --> [a].

noun --> [cat].
noun --> [cats].
noun --> [mouse].
noun --> [mice].

verb --> [scare].
verb --> [scares].
verb --> [hate].
verb --> [hates].
```

5.6 Tekstin esikäsittely

- "roskan" poistaminen
- paloittelu käsiteltäviin yksiköihin (segmentointi tai tokenisointi)
- normalisointi: variaation eliminointi, monitulkintaisuuksien ratkominen

Roskan poistaminen

Kaikki mikä ei ole varsinaista tekstiä poistetaan, esim.

- Samoina toistuvat tiedot dokumenteissa (headerit, footerit)
- Kenttien nimet
- sähköpostiviestien headerit, signaturet

Perl-esimerkki (poista tekstit, jotka ovat sulkeiden sisällä):

```
#!/usr/bin/perl
while (<STDIN>) {
   s/\(.*?\)//g;
   print $_;
}
```

Paloittelu

Mikä on sana? Kahden tyhjän tilan (blanko, space) erottama alue? Puheessa kahden tauon erottama äänisignaali? Entä

- 'database' vs. 'data base'
- yhdyssanat: kesäilta, koti-ilta, venäläissotilaat, sivukonttori, elinkeinoelämä: elin keino elämä, elinkeino elämä, elin keinoelämä?
- rikas taivutus: huomaamattomuudellaansakaankohan huomaa mattomuu dellaan sa kaan ko han ?
- "John's": 1,2 vai 3 sanaa?
- puhelinnumerot, e-mail-osoitteet
- puheessa äänisignaalin tauot eivät osu sananrajoille vaan tietyille konsonanteille

Mikä on virke? (pisteen monitulkintaisuus, muut tavat lopettaa virke)

Mikä on dokumentti? Miten käsitellään pitkät dokumentit esim. indeksoitaessa tiedonhakua varten, paloina vai kokonaisina?

5.7 Variaatio informaation koodaustavoissa

Esim. käsite ITSEORGANISOIVA KARTTA; ilmauksia SOM, SOFM, self-organizing map, self-organising map, self organizing map, self-organizing feature map, Kohonen map, Kohonen SOM, Kohonen net.

Suomeksi: itseorganisoiva kartta, itseorganisoiTUva kartta, itsejärjestyvä kartta, Kohosen kartta, Kohosen verkko

Puhelinnumerojen koodaustapoja mm.:

040 123 4567	Suomi
040-123 4567	
040-1234567	
+358-40-1234567	
(040) 123 4567	
+411/284 3797	Sveitsi
(44.171) 830 1007	UK
+44 (0) 171 830 1007	
1-925-225-3000	USA
212.995.5402	

Information extraction, informaation ekstrahointi: yritetään oppia eri tavat ilmaista sama semanttinen informaatio. Tämä on *hahmontunnistusta*, pyrkimyksenä tunnistaa tietyn semanttisen informaation tyypit.

Morfologisen monimuotoisuuden käsittely

Sanan katkaisu 'juurimuotoon' (stemming)
'istuimme', 'istuttiin' .. → 'istu'
'yöt'→ 'yö'
'öisin'→ 'öi'

sanan perusmuotoistus
etsimme' → 'etsiä'
'lying' → 'lie'/'lay'
'istui', 'istuu', 'istunut', 'istumme' → 'istua'
'istuisitko', 'istuttaisiinko', 'ISTU!' → 'istua'

Kannattaako morfologinen variaation normalisointi tehdä, tai missä määrin?

Riippuu tavoitteesta, esim. halutaanko analysoida keskusteluja yksityiskohtaisesti (esim. interaktiivinen keskustelujärjestelmä) vai representoida pääasiallise puheenaiheet (esim. tiedonhaku).

5.8 Monitulkintaisuus (ambiguity)

Samalla sanalla tai ilmauksella voi olla monta erilaista tulkintaa:

- Englannin *title*: kirjan otsikko, elokuvan nimi, henkilön nimen etuliite tai arvonimi, omistusoikeus, jne
- *Kun lakkaa satamasta, hae lakkaa satamasta.*
- *'...pääsi perille...'* Montako mahdollista tulkintaa?

Edellisen kalvon monitulkintaisuuskysymykseen liittyen
(Lingsoftin TWOL)

<*pääsi*> "pää" N NOM SG 2SG
"pää" N GEN SG 2SG
"pää" N NOM PL 2SG
"päästä" V PAST ACT SG3

<*perille*> "perä" N ALL PL
"perille" ADV ALL
"per" PROP N ALL SG

Mahdollisia tulkintoja ainakin:

- saapui sinne minne oli menossa
- saapui Per:in luo
- 'paina tämä asia pääsi perimmäiseen nurkkaan'
- 'näytitkö pääsi Perille?' (jos siinä oli vaikkapa haava ja Per on lääkäri)

Disambiguointi I. yksikäsitteistäminen

- Valitaan potentiaalisista tulkinnoista oikea tai todennäköisin
- Disambiguointi on tyypillinen kieliteknologinen tehtävä. Esim. morfologian yksikäsitteistäminen, lauserakenteen yksikäsitteistäminen, sananmerkitysten yksikäsitteistäminen
- Voidaan periaatteessa tehdä kontekstin perusteella; edellyttää *mallia* kontekstin ja vaihtoehtoisten tulkintojen välisistä riippuvuuksista.
- Tilastollinen malli: ehdollinen todennäköisyysjakauma $p(\text{tulkinta}|\text{konteksti})$
- Ei-tilastollinen malli: kategoriset säännöt muotoa 'JOS konteksti NIIN tulkinta'.

5.9 Taggaus

Syntaktisia tagijoukkoja Englannin kielelle

Brown, Penn, Claws 1-3

Taggauksen ääripäät eri tarkoituksiin

Puheentunnistus: Pelkästään puheessa esiintyvät sanat (ei välttämättä edes välimerkkejä)

Keskusteluanalyysi: vuoronvaihdot, vuorotyypit, epäröinnit, hiljaisuuden kesto jne.

Category	Examples	Claws c5	Brown	Penn
Adjective	happy, bad	AJ0	JJ	JJ
Adjective, ordinal number	sixth, 72nd, last	ORD	OD	JJ
Adjective, comparative	happier, worse	AJC	JJR	JJR
Adjective, superlative	happiest, worst	AJS	JJT	JJS
Adjective, superlative, semantically	chief, top	AJ0	JJS	JJ
Adjective, cardinal number	3, fifteen	CRD	CD	CD
Adjective, cardinal number, one	one	PNI	CD	CD
Adverb	often, particularly	AV0	RB	RB
Adverb, negative	not, n't	XX0	*	RB
Adverb, comparative	faster	AV0	RBR	RBR
Adverb, superlative	fastest	AV0	RBT	RBS
Adverb, particle	up, off, out	AVP	RP	RP
Adverb, question	when, how, why	AVQ	WRB	WRB
Adverb, degree & question	how, however	AVQ	WQL	WRB
Adverb, degree	very, so, too	AV0	QL	RB
Adverb, degree, postposed	enough, indeed	AV0	QLP	RB
Adverb, nominal	here, there, now	AV0	RN	RB
Conjunction, coordination	and, or	CJC	CC	CC
Conjunction, subordinating	although, when	CJS	CS	IN
Conjunction, complementizer *that*	that	CJT	CS	IN
Determiner	this, each, another	DT0	DT	DT
Determiner, pronoun	any, some	DT0	DTI	DT
Determiner, pronoun, plural	these, those	DT0	DTS	DT
Determiner, prequalifier	quite	DT0	ABL	PDT
Determiner, prequantifier	all, half	DT0	ABN	PDT
Determiner, pronoun or double conj.	both	DT0	ABX	DT (CC)
Determiner, pronoun or double conj.	either, neither	DT0	DTX	DT (CC)
Determiner, article	the, a, an	AT0	AT	DT
Determiner, postdeterminer	many, same	DT0	AP	JJ
Determiner, possessive	their, your	DPS	PP$	PRP$
Determiner, possessive, second	mine, yours	DPS	PP$$	PRP
Determiner, question	which, whatever	DT0	WDT	WDT

6. Kollokaatiot

6.1 Mitä on kollokaatio

- Kahdesta tai useammasta sanasta koostuva konventionaalistunut ilmaus (Manning & Schütze)
- Collocations of a given word are statements of the habitual or customary places of that word (Firth, 1957)
- Esimerkkejä:
  - 'weapons of mass destruction', 'disk drive', 'part of speech' (suomessa yhdyssanoina 'joukkotuhoaseet', 'levyasema', 'sanaluokkatieto')
  - 'bacon and eggs'
  - verbin valinta: 'prendre une décision', mutta 'make a decision' ei 'take a decision'.

  - adjektiivin valinta: 'strong tea' mutta ei 'powerful tea'; 'vahvaa teetä', harvemmin 'voimakasta teetä' (valinnat voivat heijastaa kulttuurin asenteita: strong → tea, coffee, cigarettes powerful → drugs, antidote)
  - 'kick the bucket', 'heittää veivinsä' (kiertoilmaus, sanonta, idiomi)

- Olentoja, yhteisöjä, paikkoja tai tapahtumia yksilöivät nimet: 'White House' Valkoinen talo, 'Tarja Halonen', 'Persianlahden sota' (viittaa tiettynä ajankohtana käytyyn sotaan)

- Kollokaation kanssa osittain päällekkäisiä käsitteitä: termi, tekninen termi, terminologinen fraasi. Huom: tiedonhaussa sanalla 'termi' laajempi merkitys: 'sana tai kollokaatio'.

Kollokaatioiden ominaisuuksia

- Ei-kompositionaalisuus: kollokaation merkitys ei ole (täysin) selitettävissä osiensa summana
- Ei-vaihdettavuus: 'white wine' mutta ei 'yellow wine' ('valkoviini', ei 'keltainen viini', todellisesta väristä huolimatta)
- (Ei-muutettavuus: 'Ihmiset seurasivat tilannetta *pala kurkussa*', ei 'pieni pala kurkussa' tai 'palat kurkuissa'. 'Pitäkää päät pystyssä' vs. 'pitäkää pää pystyssä'.)

6.2 Sovelluskohteita

- Luonnollisen kielen generointi: vältetään kummallisia sanavalintoja.
- Lauseenjäsennys: suositaan konventionaalisia sanayhdistelmiä rakenteellisina yksiköinä rakenteen disambiguoinnissa.
- Leksikografia (sanakirjojen tekeminen): mitkä ilmaukset valitaan sanakirjassa selitettäviin.
- Tiedonhaku ja indeksointi: sanaparien valinta indeksointitermeiksi.
- Tietokoneavusteinen kielenkääntäminen: tekniset termit käännettävä systemaattisesti samaa ilmausta käyttäen.
- Korpuslingvistiikka: kulttuuristen asenteiden ja arvostusten tutkiminen, mm. eri nautintoaineita tai eri sukupuolia kohtaan.