5. Kontekstitieto ja yhteisesiintyminen

- Kontekstin tärkeys kielen tulkinnassa: esimerkiksi monitulkintaisuudet (“Aloitin alusta”, “Alusta kovalevy!”, “Näin monta alusta”, “Minä näin monta alusta”)
- Chomskyn hierarkia kielille
- Kontekstin pituus
- Yhteisesiintymismatriisi
- Kollokaatiot

5.1 Chomskyn hierarkia kielille

Chomsky jakaa kielet seuraavanlaiseen kompleksisuushierarkiaan.

Tässä A on yksittäinen ei-terminaalisymboli ja α, β ja γ ovat mitä tahansa terminaalien ja ei-terminaalien jonoja.

Tyyppi	**Nimi**	**Sääntörunko**
0	Turing-ekvivalentti	$\alpha \rightarrow \beta$ s.e. $\alpha \neq \epsilon$
1	Kontekstiherkkä	$\alpha A \beta \rightarrow \alpha \gamma \beta$ s.e. $\gamma \neq \epsilon$
2	Kontekstivapaa	$A \rightarrow \gamma$
3	Säännöllinen	$A \rightarrow xB$ tai $A \rightarrow x$

Chomskyn hierarkia kielille, selitykset

Tyyppi 0 vastaavat niitä kieliä, joiden symbolijonot voidaan tuottaa (listata) Turing-koneella.

Kontekstiherkät kielet muuntavat symbolin toiseksi riippuen sen oikean- ja vasemmanpuoleisesta kontekstista. Sääntöjen on myös tuotettava jotakin.

Kontekstivapaissa kielissä ei-terminaalisymboli voidaan korvata millä tahansa ei-terminaalien ja terminaalien jonolla (mukaanlukien tyhjä jono). Näitä toteuttavat esim. lauserakennekieliopit.

Säännölliset kielet ovat ekvivalentteja säännöllisten lausekkeiden (regular expressions) kanssa. Ne voivat olla oikea- tai vasenkätisesti lineaarisia. Niitä toteuttavat esim. äärelliset tilakoneet.

5.2 Kontekstin pituus ja yhteisesiintymismatriisi

- n-grammit, dynaaminen konteksti
- yksittäiset sanat kontekstissa, sana-dokumenttimatriisi
- bag of words -malli versus sanapositioiden huomioiminen
- etäriippuvuudet ja lauserakenteen huomioiminen

5.3 Kollokaatiot

- Kollokaatio on kahdesta tai useammasta sanasta koostuva konventionaalistunut ilmaus
- Esimerkkejä:
  - 'weapons of mass destruction', 'disk drive', 'part of speech' (suomessa yhdyssanoina 'joukkotuhoaseet', 'levyasema', 'sanaluokkatieto')
  - 'bacon and eggs'
  - verbin valinta: '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 $\rightarrow$ tea, coffee, cigarettes powerful $\rightarrow$ drugs, antidote)
  - 'kick the bucket', 'heittää veivinsä' (kiertoilmaus, sanonta, idiomi)

- Olentoja, yhteisöjä, paikkoja tai tapahtumia yksilöivät nimet: 'White House' Valkoinen talo, 'Tarja Halonen'

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

Sanan frekvenssi ja sanaluokkasuodatus

Pelkän frekvenssin käyttö:

Esimerkki: Onko luontevampaa sanoa 'strong tea' vai 'powerful tea'?
Ratkaisu: Etsitään Googlella: 'strong tea' 9270, 'powerful tea' 201

Joihinkin täsmällisiin kysymyksiin riittävä tapa. Kuitenkin järjestettäess bigrammeja frekvenssin mukaan, parhaita ovat 'of the', 'in the', 'to the', ...

Frekvenssi + sanaluokka:

Jos tunnetaan kunkin sanan sanaluokka, sekä osataan kuvailla kollokaatioiden 'sallitut' sanaluokkahahmot:

- Järjestetään sanaparit tai -kolmikot yleisyyden (lukumäärä) mukaan
- Hyväksytään vain tietyt sanaluokkahahmot:
  AN, NN, AAN, ANN, NAN, NNN, NPN (Justeson & Katz's POS filter)

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Sanojen etäisyyden keskiarvo ja varianssi

Entä joustavammat kollokaatiot, joiden keskellä on kollokaatioon kuulumattomia sanoja?

Lasketaan etäisyyden keskiarvo ja varianssi. Jos keskiarvo nollasta poikkeava ja varianssi pieni, potentiaalinen kollokaatio (Huom: oletetaan siis etäisyyden jakautuvan gaussisesti).

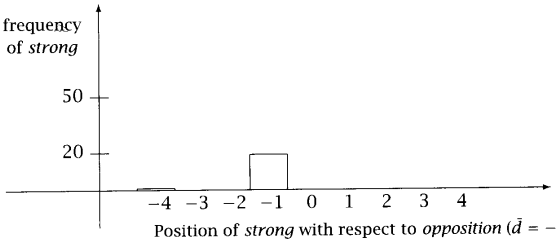
Esim. '*knock . . . door*' (ei 'hit', 'beat', tai 'rap'):
a) 'She *knocked* on his *door*'
b) 'They *knocked* at the *door*'
c) '100 women *knocked* on Donaldson's *door*'
d) 'a man *knocked* on the metal front *door*'

Algoritmi

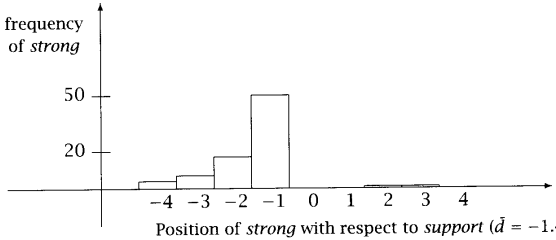
- Liu'uta kiinteän kokoista ikkunaa tekstin yli (leveys esim. 9) ja kerää kaikki sanaparin esiintymät koko tekstissä

- Laske sanojen etäisyyksien keskiarvo:
  $\bar{d} = 1/n \sum_{i=1}^{n} d_i = 1/4(3+3+5+5) = 4.0$
  (jos heittomerkki ja 's' lasketaan sanoiksi)

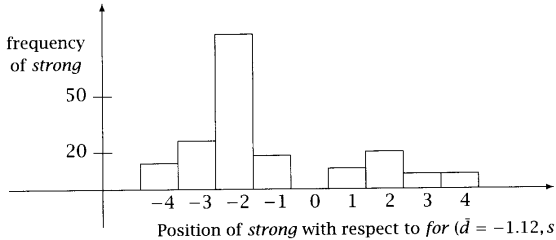
- Estimoi varianssi s^2 (pienillä näytemäärillä):
  $s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = 1/3((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)$
  $s = 1.15$



Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Table 5.5 Finding collocations based on mean and variance. Sample deviation s and sample mean $\bar{d}$ of the distances between 12 word pairs.

Pohdittavaksi:

1. Mitä tapahtuu jos sanoilla on kaksi tai useampia tyypillisiä positioita suhteessa toisiinsa?
2. Mikä merkitys on ikkunan leveydellä?

Hypoteesin testaus

Onko suuri osumamäärä yhteensattumaa (esim. johtuen siitä että jommankumman perusfrekvenssi on suuri)? Osuvatko kaksi sanaa yhteen useammin kuin sattuma antaisi olettaa?

1. Formuloi *nollahypoteesi* H_0: assosiaatio on sattumaa
2. Laske tn p että sanat esiintyvät yhdessä jos H_0 on tosi
3. Hylkää H_0 jos p liian alhainen, alle merkitsevyystason, esim $p < 0.05$ tai $p < 0.01$.

Nollahypoteesia varten sovelletaan riippumattomuuden määritelmää.

Oletetaan että sanaparin todennäköisyys, jos H_0 on tosi, on kummankin sanan oman todennäköisyyden tulo:
$P(w^1 w^2) = P(w^1)P(w^2)$

T-testi

Tilastollinen testi sille eroaako havaintojoukon odotusarvo oletetun, datan generoineen jakauman odotusarvosta. Olettaa, että todennäköisyydet ovat suunnilleen normaalijakautuneita.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}, \text{ jossa} \tag{1}$$

$\bar{x}, s^2$: näytejoukon keskiarvo ja varianssi, $N =$ näytteiden lukumäärä, ja $\mu =$ jakauman keskiarvo. Valitaan haluttu p-taso (0.05 tai pienempi). Luetaan tätä vastaava t:n yläraja taulukosta. Jos t suurempi, H_0 hylätään.

Soveltaminen kollokaatioihin:

Nollahypoteesina että sanojen yhteisosumat ovat satunnaisia: Esimerkki: H_0 : $P(new\ companies) = P(new)P(companies)$

$\mu = P(new)P(companies)$
$\bar{x} = \frac{c(new\ companies)}{c(\cdot,\cdot)} = \hat{p}$
$s^2 = p(1-p) = \hat{p}(1-\hat{p}) \approx \hat{p}$ (pätee Bernoulli-jakaumalle)
$N = c(\cdot,\cdot)$

- Järjestetään sanat paremmuusjärjestykseen mitan mielessä TAI
- Hypoteesin testaus: valitaan merkittävyystaso (p=0.05 tai p=0.01) ja katsotaan t-testin taulukosta arvo, jonka ylittäminen tarkoittaa nollahypoteesin hylkäystä.

Vertaillaan yhtä suuren frekvenssin omaavia bigrammeja keskenään t-testillä:

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Table 5.6 Finding collocations: The t test applied to 10 bigrams that occur with frequency 20.

Esimerkki soveltamisesta muuhun ongelmaan: Vertailu mitkä lähikonteks sanat parhaiten erottelevat sanoja 'strong' ja 'powerful'

t	$C(w)$	$C(strong\ w)$	$C(powerful\ w)$	Word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition
3.9000	802	18	1	showing
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

Table 5.7 Words that occur significantly more often with *powerful* (the first ten words) and *strong* (the last ten words).