Harjoitustehtävät

1. Arvioi n-grammitodennäköisyydet (n=1,2,3) korpuksesta

adddd
bcbcebc
bcbcedd
bcdddbc
dadeae

kun jokaista korpuksen riviä käsitellään erillisenä lauseena.

Laadi n-grammeja järkevällä katsomallasi tavalla approksimoivat probabilistiset yhteysvapaat kieliopit, jotka täyttävät ehdot:

1. Aloitussymboli on S
2. terminaalisymbolit ovat {a,b,c,d,e}
3. välisymbolit ovat A_i (i=1,...,N)
4. Kieliopissa on säännöt

$A_1 \rightarrow a\ (p=1)$ $A_2 \rightarrow b(1)$

$A_3 \rightarrow c(1)$ $A_4 \rightarrow d(1)$

$A_5 \rightarrow e(1)$

5. Lisäksi voi olla sääntöjä

$S -> A_i \quad A_i \rightarrow e$

$A_i \rightarrow A_j A_k$ (i ≠ j ≠ k)

$A_i \rightarrow A_j \mid A_k \; (p = 0{,}5;0{,}5)$

Sääntötodennäköisyydet, joita ei ole annettu, saa vapaasti valita.

Jos keksit hyvän paramatrisoidun esitystavan säännöille, saa sitä käyttää eikä kaikkia sääntöjä tarvitse luetella erikseen. Aproksimaation järkevyyttä arvioidessa voi olettaa, että kielioppia käytetään tehtävän 2 kaltaisiin tarkoituksiin.

Tehtävä 2.

Määrittele järkevä koodaus 1. tehtävän kieliopeille. Laske kullekin kieliopille korpuksen ja kieliopin yhteenlaskettu koodinpituus.