

Chomskyn hierarkia ja yhteysherkät kieliopit

Tekijät:

Teemu Hurme

Kati Kyllönen

Sisällysluettelo

1. Chomskyn hierarkia.....	3
1.1 Chomskyn hierarkian laatijat.....	3
1.2 Chomskyn hierarkian kieliopit.....	3
Kuva 1: Chomskyn hierarkia.....	4
1.3.1 Termistöä.....	4
2. Rajoittamattomat kieliopit.....	5
3. Yhteysherkät kieliopit.....	5
Kuva 2: Lineaarisesti rajoitettu Turingin kone.....	7
4. Yhteysvapaat kieliopit.....	7
5. Säännölliset kieliopit.....	8
6. Yhteenveto.....	9
Lähteet:	9

1. Chomskyn hierarkia

1.1 Chomskyn hierarkian laatijat

Chomskyn hierarkia on saanut nimensä kielitieteen emeritusprofessorilta Noam Chomskyltä, joka laati hierarkian yhdessä matemaatikko Marcel-Paul Schützenbergerin kanssa vuonna 1956. Chomsky tutki, miten luonnollisia kieliä voi kuvata muodollisilla kieliopeilla ja laati teorian generoivista kieliopista. Kielioppien rajoitusten ja niiden ilmaisuvoiman keskinäisten suhteiden tutkiminen johti Chomskyn hierarkian syntymiseen. Chomskyn kielitieteelliset tutkimustulokset ovat merkittäviä tietojenkäsittelytieteessä.

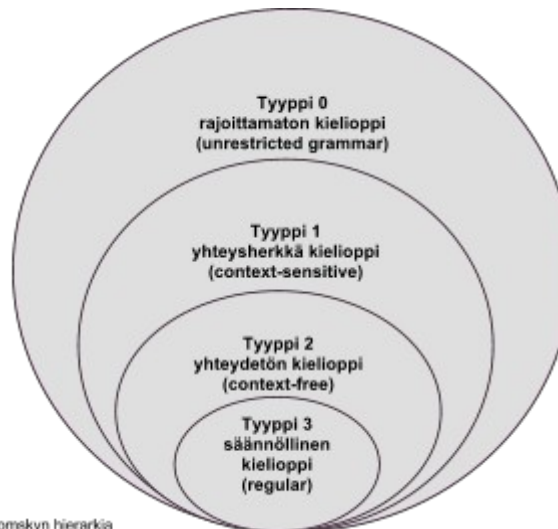
1.2 Chomskyn hierarkian kieliopit

Chomskyn hierarkiaan kuuluvat kieliopit ovat generoivia, jolla tarkoitetaan sitä, että kieliopin sääntöjen avulla voidaan tuottaa kaikki kieleen kuuluvat merkkijonot. Kielioppi määritellään nelikkona $G = (V, \Sigma, R, S)$. V ja Σ ovat muuttuja- ja päätesymbolien joukkoja, R on korvaussääntöjen joukko ja S on aloitussymboli. Kieliopin määrittelyssä käytetyt joukot ovat kooltaan äärellisiä, mutta kieliopin tuottamaan kieleen voi kuulua rajattomasti erilaisia merkkijonoja.

Kieliopin voidaan ajatella olevan kokoelma korvaussääntöjä, joita käyttämällä saadaan tuotettua toisia merkkijonoja, jotka voivat sisältää muuttuja ja päätesymboleita. Tuotettuihin merkkijonoihin voidaan soveltaa korvaussääntöjä uudestaan. Jos lopputuloksena saatuun merkkijonoon kuuluu ainoastaan päätesymboleita, kuuluu se kyseiseen kieleen. Korvaussääntöjen soveltamisjärjestystä, jonkin tietyn merkkijonon tuottamiseksi, kutsutaan johdoksi. On mahdollista, että saman merkkijonon voi tuottaa usealla eri johdolla.

Kieliopin korvaussäännöt ovat muotoa $\alpha \rightarrow \beta$, missä merkkijonojen α on korvattava muuttujasymboli tai merkkijono ja β on tuotettava merkkijono. Jäljempänä käytämme α :sta myös nimitystä säännön vasen puoli ja β :stä nimitystä säännön oikea puoli.

Chomskyn hierarkia luokittelee kieliopit neljään tyyppiin, jotka ovat: rajoittamaton kielioppi, yhteysherkkä kielioppi, yhteysvapaa kielioppi ja säännöllinen kielioppi. Seuraavan luokan kielioppi on ilmaisuvoimaltaan aina edellistä heikempi, johtuen korvaussäännöille asetetuista lisärajoituksista.



Kuva 1: Chomskyn hierarkia

1.3.1 Termistöä

$G=(V, \Sigma, R, S)$	kieliopin määritelmä
V	äärellinen muuttujasymbolien joukko
Σ	äärellinen päätesymbolien joukko, joka ei saa sisältää muuttujasymboleita
R	äärellinen joukko korvaussääntöjä
S	aloitussymboli ($S \in V$)
ε	tyhjä merkki
L	kieli
w, α, β	Merkkijono

2. Rajoittamattomat kieliopit

Chomskyn hierarkian ilmaisuvoimaisinta kielioppia kutsutaan rajoittamattomaksi kieliopiksi (Unrestricted Grammar).

Rajoittamattoman kieliopin säännöt ovat muotoa $\alpha \rightarrow \beta$, missä α ja β ovat merkkijonoja. Nämä merkkijonot koostuvat muuttuja- ja päätesymboleista. Rajoittamattoman kieliopin korvaussääntöjen ainoa rajoitus on, että säännön vasemmanpuoleinen osa ei saa olla tyhjä merkkijono, eli $\alpha \neq \varepsilon$.

Kielioppi kielelle $L = \{ a^i \mid i = 2, 4, 8, \dots \}$

- | | |
|-------------------------|---------------------------------|
| 1. $S \rightarrow ACab$ | 5. $aD \rightarrow Da$ |
| 2. $S \rightarrow aaC$ | 6. $AD \rightarrow AC$ |
| 3. $CB \rightarrow DB$ | 7. $aE \rightarrow Ea$ |
| 4. $CB \rightarrow E$ | 8. $AE \rightarrow \varepsilon$ |

Johto kieleen L kuuluvalla merkkijonolla aaaa

$S \rightarrow ACaB \rightarrow AaaCB \rightarrow AaaDB \rightarrow AaDaB \rightarrow ADaaB \rightarrow AaaCaB \rightarrow AaaaaCB \rightarrow AaaaaCB$
 $\rightarrow AaaaaE \rightarrow AaaaEa \rightarrow AaaEaa \rightarrow AaEaaa \rightarrow AEaaaa \rightarrow aaaa$

Rajoittamaton kielioppi ja Turingin kone ovat ekvivalentteja, sillä kaikille rajoittamattomille kieliopille voidaan muodostaa Turingin kone, joka tunnistaa kaikki kieleen kuuluvat merkkijonot. Vastaavasti jokaiselle Turingin koneen tunnistamalle kielelle voidaan muodostaa tämän kielen tuottava rajoittamaton kielioppi [HoF69/1].

3. Yhteysherkät kieliopit

Yhteysherkkä kielioppi (Context-sensitive Grammar) saadaan muodostettua lisäämällä rajoittamattoman kieliopin korvaussääntöihin seuraava rajoitus: korvaussäännön avulla muodostetun merkkijonon pituuden on oltava vähintään yhtä pitkä kuin säännön vasen puoli, eli $\alpha \leq \beta$.

Nimi yhteysherkkä tulee kieliopin normaalimuodosta, jossa sääntöjen on oltava muotoa $\alpha A \beta \rightarrow \alpha w \beta$, missä α ja β ovat yhteys joka säilyy muunnoksessa. Muuttajasymboli A korvataan merkkijonolla w, joka voi sisältää sekä muuttuja- että päätesymboleja.

Kielioppi kielelle $L = \{a^n b^n c^n\}$

1. $S \rightarrow abc | aSQ$
2. $bQc \rightarrow bbcc$
3. $cQ \rightarrow Qc$

Johto kieleen L kuuluvalla merkkijonolla aaabbbccc

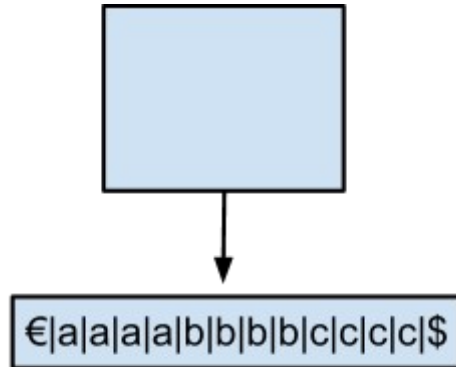
$S \rightarrow aSQ \rightarrow aaSQQ \rightarrow aaabcQQ \rightarrow aaabQcQ \rightarrow aaabbccQ \rightarrow aaabbcQc \rightarrow aaabbQcc \rightarrow aaabbbccc$

Rajoittamattomien kielioppien yhteydessä esittelemämme kielioppi kielelle $L = \{a^i \mid i=2,4,8,\dots\}$ voidaan muuntaa yhteysherkäksi kieliopiksi [HoF69/2].

Yhteysherkän kieliopin tuottamat merkkijonot voidaan hyväksyä lineaarisesti rajoitetulla Turingin koneella. Lineaarisesti rajoitettu Turingin kone (Linear Bounded Automaton) on Turingin kone, jonka käyttämän nauhan pituus on rajoitettu syötteen mittaiseksi. Nauhalle kirjoitetaan syötteen lisäksi alku- ja loppumerkit, joista lineaarisesti rajoitettu Turingin kone tunnistaa käytettävissä olevan nauhan pituuden. Nauhan alku- ja loppumerkit rajoittavat lukupään liikkumista, eikä niiden päälle saa kirjoittaa. Jos lukupäätä yritetään siirtää nauhan alku- tai loppumerkin ohi, lukupää pysyy paikoillaan. Lineaarisesti rajoitetun Turingin koneen käytettävissä olevan nauhan pituutta saadaan lisättyä laajentamalla nauha-aakkostoa. Tällöin jokainen alkuperäisen syötteen merkki voidaan koodata nauhalle useammalla merkillä. Alkuperäiseen nauha-aakkostoon kuulumattomat merkit voidaan myöhemmin poistaa, jolloin vapautuva tila nauhalla jää koneen lisämuistiksi. Näin käytettävissä olevan nauhan pituutta saadaan lisättyä jollakin vakiokertoimella, mistä johtuu nimi lineaarisesti rajoitettu. Nauhan pituuden kasvattaminen ei kuitenkaan vaikuta koneen laskentatehoon [HoF69/3].

Nauhan rajallisesta pituudesta seuraa, että Turingin koneen pysähtymisongelma on ratkeava lineaarisesti rajoitetulla Turingin koneella, sillä mahdollisten tilojen ja nauhan konfiguraatioiden määrä on rajallinen [Sip06/2].

Yhteysherkkä kieli voidaan hyväksyä lineaarisesti rajoitetulla Turingin koneella siten, että kone valitsee epädeterminisesti kohdan syötteenä saamastaan merkkijonosta ja muuntaa tämän osamerkkijonon käyttämällä epädeterminisesti valittua korvaussääntöä oikealta vasemmalle. Koska merkkijono tällöin lyhenee tai pysyy saman mittaisena, riittää rajoitettu nauhan pituus kielen tunnistamiseen.



Kuva 2: Lineaarisesti rajoitettu Turingin kone

4. Yhteysvapaat kieliopit

Yhteysvapaa kieliopin (Context-free Grammar) korvaussääntöjä rajoitetaan siten, että säännön vasemmalla puolella saa olla vain yksi muuttujasymboli. Korvaussääntöä voidaan käyttää huomioimatta muuttujasymbolin ympärillä olevia merkkejä, eli tällöin korvaussäännön katsotaan olevan yhteysvapaa. Lisäksi yhteysvapaa kielioppi sallii säännöt, jotka muuttavat muuttujasymbolin tyhjäksi merkiksi.

Yhteysvapaiden kielten pumppauslemman avulla voidaan todistaa, jos kieli ei kuulu yhteysvapaiden kielten luokkaan. Pumppauslemman avulla ei kuitenkaan voida todistaa, että kyseinen kieli kuuluisi yhteysvapaisiin kieliin, vaikka se täyttäisikin pumppauslemman ehdot.

Esimerkiksi yhteysherkkien kielioppien yhteydessä esittelemämme kieli $L = \{a^n b^n c^n\}$, voidaan osoittaa pumppauslemman avulla kuulumattomaksi yhteysvapaiden kielten luokkaan. [Sip06/1]

Kielioppi kielelle $L = \{a^n b^n c^n\}$

1. $S \rightarrow AT$
2. $A \rightarrow aA | \epsilon$
3. $T \rightarrow bTc | \epsilon$

Johto merkkijonolle aabbbccc

$S \rightarrow AT \rightarrow aAT \rightarrow aaAT \rightarrow aaT \rightarrow aabTc \rightarrow aabbTcc \rightarrow aabbbTccc \rightarrow aabbbccc$

Yhteysvapaa kieliopin tuottaman merkkijonon voi tunnistaa pinoautomaatin avulla.

5. Säännölliset kieliopit

Chomskyn hierarkian rajoitetuin kielioppi on säännöllinen kielioppi (Regular Grammar). Säännöllinen kielioppi rajoittaa korvaussäännöt, joko muotoon $A \rightarrow w\beta$, jolloin kyseessä on oikealle lineaarinen kielioppi, tai muotoon $A \rightarrow \beta w$, jolloin kyseessä on vasemmalle lineaarinen kielioppi. Tässä A voi olla mikä tahansa kielioppiin kuuluva muuttujasymboli, w on aiemmin tuotettu merkkijono joka sisältää ainoastaan päätesymboleita ja β on kyseisellä säännöllä muodostettu merkkijono.

Oikealle lineaarinen kielioppi kielelle $L = \{ 0(10)^n \mid n \geq 0 \}$

1. $S \rightarrow 0A$
2. $A \rightarrow 10A \mid \varepsilon$

Johto $S \rightarrow 0A \rightarrow 010A \rightarrow 01010A \rightarrow 0101010A \rightarrow 0101010$

Vastaava vasemmalle lineaarinen kielioppi kielelle $L = \{ 0(10)^n \mid n \geq 0 \}$

1. $S \rightarrow S10 \mid 0$

Säännöllisen kieliopin tuottaman merkkijonon voi tunnistaa äärellisellä automaatilla. Kaikille säännöllisille kieliueille löytyy vastaava säännöllinen lauseke.

6. Yhteenveto

Oheisessa taulukossa on esitetty tiivistettynä Chomskyn hierarkian kielet.

Tyyppi	Kielioppi	Sääntöesimerkki	Kone	Sulkeumat [HoF69/4]
0	Rajoittamaton	$Ca \rightarrow aaC$	Turingin kone	yhdiste, leikkaus, Kleenen tähti, kääntö, konkatenaatio
1	Yhteysherkkä	$bQc \rightarrow bbcc$	lineaarisesti rajoitettu Turingin kone	yhdiste, leikkaus, Kleenen tähti, kääntö, konkatenaatio, komplementti
2	Yhteysvapaa	$T \rightarrow bTc \mid \varepsilon$	pinoautomaatti	yhdiste, Kleenen tähti, kääntö, konkatenaatio
3	Säännöllinen	$A \rightarrow 10A \mid \varepsilon$	äärellinen automaatti	yhdiste, leikkaus, Kleenen tähti, kääntö, konkatenaatio, komplementti

Lähteet:

[HoF69/1] J.E. Hopcroft, J.D. Ullman. Formal Languages and their relation to automata. Addison-Wesley, 1969, sivu 222

[HoF69/2] J.E. Hopcroft, J.D. Ullman. Formal Languages and their relation to automata. Addison-Wesley, 1969, sivu 224

[HoF69/3] J.E. Hopcroft, J.D. Ullman. Formal Languages and their relation to automata. Addison-Wesley, 1969, sivu 225

[HoF69/4] J.E. Hopcroft, J.D. Ullman. Formal Languages and their relation to automata. Addison-Wesley, 1969, sivu 280-281

[Sip06/1] M.Sipser. Introduction to the Theory of Computation. Thomson Course Technology, 2006, sivu 128

[Sip06/2] M.Sipser. Introduction to the Theory of Computation. Thomson Course Technology, 2006, sivu 198