

hyväksymispäivä

arvosana

arvostelija

Historiakatsaus assemblykääntäjistä korkean tason kielten kääntäjiin

Arttu Kilpinen

Helsinki 5.12.2016

HELSINGIN YLIOPISTO

Tietojenkäsittelytieteen laitos

Tiedekunta — Fakultet — Faculty		Laitos — Institution — Department	
Matemaattis-luonnontieteellinen tiedekunta		Tietojenkäsittelytieteen laitos	
Tekijä — Författare — Author			
Arttu Kilpinen			
Työn nimi — Arbetets titel — Title			
Historiakatsaus assemblykääntäjistä korkean tason kielten kääntäjiin			
Oppiaine — Läroämne — Subject			
Tietojenkäsittelytiede			
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages	
	5.12.2016	21 sivua	
Tiivistelmä — Referat — Abstract			
<p>Ensimmäiset ohjelmointikielten kääntäjät, assemblykääntäjät, käänsivät symbolisille konekielille kirjoitettuja ohjelmia konekielisiksi ohjelmiksi. Täsmällisten kuvausjärjestelmien kehitys sekä koodin generoinnin teoria mahdollistivat tehokkaampien ohjelmointikielten kehityksen. Nykyisin käytössä olevat korkean tason ohjelmointikielet kehittyivät hiljalleen kuvausjärjestelmien kehittyessä ja syrjäyttivät symbolisella konekielillä ohjelmoinnin lähes kokonaan. Tässä dokumentissa käydään läpi historiallisia vaiheita symbolisten konekielten kääntäjistä nykyaikaisiin korkean tason ohjelmointikielten kääntäjiin. Läpi käydään useita merkittäviä ohjelmointikieliä ja niiden ominaisuuksia.</p> <p>ACM Computing Classification System (CCS): Software and its engineering -> software notations and tools -> Compilers</p>			
Avainsanat — Nyckelord — Keywords			
Historia, Kääntäjät, Symbolinen konekieli, Ohjelmointikielet			
Säilytyspaikka — Förvaringsställe — Where deposited			
Muita tietoja — Övriga uppgifter — Additional information			

Sisältö

1	Johdanto	1
2	Symbolinen konekieli ja assemblykääntäjät	2
2.1	Historia ensimmäisistä assemblykääntäjistä	2
2.2	Assemblykääntäjien toiminnasta ja toteutuksesta	3
3	Historiaa korkean tason kielten kääntäjistä	4
3.1	Täsmällisten kuvausjärjestelmien kehitys	4
3.2	Kohti ensimmäisiä kääntäjiä	5
4	Välikielet ja ristiinkääntäminen	7
4.1	Ristiinkääntäminen	8
4.2	Välikielet	8
5	T-kaaviot	10
5.1	Alkuperäinen Bratman-kaavio	10
5.2	Earleyn ja Sturginsin merkintätapa	11
6	Kääntäjien rakenne ja ketjutus	13
6.1	Iteratiivinen ketjutus	13
6.2	Ketjutus olemassa olevien kääntäjien avulla	14
6.3	Ketjuttamista helpottava kääntäjä rakenne	16
7	Yhteenveto	18

1 Johdanto

Kääntäjät ovat tietokoneohjelmia, jotka kääntävät lähdekielisen ohjelmakoodin kohdekieliseksi ohjelmaksi (Bauer, 1974: 1). Kohdekielenä on usein jonkin prosessoriarkkitehtuurin ymmärtämä konekieli.

Ohjelmointikielet sekä niitä ymmärtävät kääntäjät ja tulkit ovat keskeisessä asemassa ohjelmistotuotannossa. Kääntäjät mahdollistavat ohjelmien kirjoittamisen korkean tason ohjelmointikielillä sekä symbolisilla konekielillä, jotka puolestaan helpottavat ja nopeuttavat ohjelmointia. Niiden käyttäminen tekee ohjelmakoodista myös ymmärrettävämpää ja helpompilukuista. Yleisesti ottaen ohjelmointikielen ymmärrettävyys kasvaa abstraktiotason kasvaessa. Esimerkiksi matemaattisesti tutulla tavalla kirjoitetut aritmeettiset lausekkeet ovat ymmärrettävämpiä kuin vastaava laskeenta symbolisella konekielellä ohjelmoituna. Lisäksi useat korkean tason ohjelmointikielet mahdollistavat — mikäli tarvittavat kääntäjät ovat olemassa — saman ohjelmakoodin käyttämisen useissa eri laitteistoissa sekä useilla eri käyttöjärjestelmillä. Koska eri laitteistoissa on erilaiset käskykannat, poistuu korkean tason ohjelmointikieliä käyttämällä myös tarve uudelleenohjelmoinnille.

Symboliset konekielet sekä korkean tason ohjelmointikielet ovat lähes yhtä vanhoja kuin ohjelmointikin. Ensimmäiset korkean tason ohjelmointikielten kääntäjät puolestaan ovat ohjelmointikieliä huomattavasti nuorempia, sillä ohjelmointikielten teoriaa kehitettiin vuosia ennen kuin ensimmäiset kääntäjät valmistuivat. Esimerkiksi ensimmäisenä korkean tason ohjelmointikielenä pidetty Plankalkül kehitettiin jo vuonna 1945, mutta sitä ymmärtävä kääntäjä valmistui vasta vuonna 1972 (Knuth and Pardo, 1976: ???).

Ennen korkeatasoisille lausekielille kehitettyjä kääntäjiä oli pitkään käytössä vain symbolisia konekieliä ymmärtäviä ohjelmia, assemblykääntäjiä. Korkean tason ohjelmointikielten kehityttyä saatiin myös niitä tukevia kääntäjiä valmistettua. Vuonna 1952 valmistunut AUTOCODEn kääntäjä oli yksi ensimmäisiä kaupallisessa ohjelmistotuotannossa käytettyjä korkean tason kielen kääntäjiä (Knuth and Pardo, 1976: 42).

2 Symbolinen konekieli ja assemblykääntäjät

Samoille asioille on tietojenkäsittelytieteessä annettu hieman toisistaan poikkeavia määritelmiä. Erään määritelmän mukaan assemblykääntäjä on kääntäjä, joka kääntää yksi yhteen symbolisella konekielellä kirjoitettuja komentoja konekielisiksi komennoiksi (Salomon, 1993: 1). Koska jokaisella laitteistolla on oma konekielensä ja tämä konekieli on myös ohjelmointikieli, pätee yleinen kääntäjien määritelmä myös assemblykääntäjiin. Lähdekielenä assemblykääntäjän ymmärtämä symbolinen konekieli tarkoittaa konekieltä, jossa laitteen ymmärtämät binääriset konekäskyt on korvattu ihmisille helpommin muistettavilla sanoilla eli symboleilla.

2.1 Historia ensimmäisistä assemblykääntäjistä

Koska ennen ensimmäisiä assemblykääntäjiä ei ollut mitään ohjelmointia helpottavia työkaluja, tuli ensimmäiset kääntäjät ohjelmoida suoraan konekielellä kuten yksi ensimmäisiä, vuonna 1949 valmistunut EDSAC tietokoneen assemblykääntäjä toteutettiin. Vaikka korkean tason kielten kääntäjät alkoivat kehittyä lähes heti ensimmäisten assemblykääntäjien valmistuttua, pysyi symbolisilla konekielillä ohjelmointi pitkään suosiossa. Alkuun korkean tason kielten automaattista käännöstyötä pidettiin lähinnä teoreettisena tutkimisena ja käytännössä kaikki ohjelmoijat uskoivat ettei automaattisesta koodin generoinnista tule ikinä tarpeeksi tehokasta oikeaan ohjelmointiin (Knuth and Pardo, 1976: 61).

Vaikka nykyaikaiset korkean tason ohjelmointikielten kääntäjät tuottavat hyvin optimoitua koodia, on hyvän ohjelmoijan kirjoittama symbolinen konekieli silti lähes poikkeuksetta parempaa. Tämän takia symbolisia konekieliä käytetään jonkin verran matalan tason ohjelmoinnin lisäksi suurta laskentatehoa vaativien ohjelmien optimointiin. Ennen symbolisten konekielten kehitystä ohjelmointi tapahtui kirjoittamalla laitteistoriippuvaista tietyn prosessorin ymmärtämää binäärikoodia. Siitä huolimatta, että käskykannat olivat nykyiseen verrattuna suhteellisen yksinkertaisia, oli ohjelmointi hidasta ja työlästä. Tietokoneiden kehittyessä ja ohjelmien monimutkaisuudessa tarve ohjelmointikielille kasvoi. Symboliset konekielet kehitettiin varhain ja nykyisin lähes kaikki sovellusohjelmat kirjoitetaan korkean tason ohjelmointikielillä.

2.2 Assemblykääntäjien toiminnasta ja toteutuksesta

Assemblykääntäjät ymmärtävät jotakin symbolista konekieltä ja osaavat tuottaa tästä konekielisen suoritettavan ohjelman. Symboliset konekielet ovat matalan tason laiteriippuvaisia ohjelmointikieliä, jotka kääntävät lähdekoodia yksinkertaisin ennalta määrättyin ehdoin kohdekielelle. Suurin osa ohjelmakoodista on siis käännettävissä yksi yhteen laitteiston ymmärtämän konekielen kanssa. Poikkeuksena on kuitenkin ohjelman osoitteina käytettävät tunnuksot (label), joiden arvot assemblykääntäjä voi vapaasti päättää. Tunnuksina ovat joko paikat ohjelman koodiosassa tai muuttujina käytetyt muistipaikat. Symbolisen konekielen avainsanat ovat siis symboleja laitteiston ymmärtämälle konekielelle. Konekielellä on mahdollista kirjoittaa suoraan suorittimen rekistereihin. Tämä tekee symbolisilla konekielillä ohjelmoimisesta yhtä laiteläheistä kuin suoraan konekielillä ohjelmointikin. Laiteläheisyys puolestaan tekee ohjelmista laitteistoriippuvaisia, sillä eri suorittimilla voi olla erilaiset käskykannat. Symbolien käyttäminen vähentää huomattavasti kirjoitusvirheiden määrää ja tekee koodista helpomman kirjoittaa ja lukea. Tunnisteiden käyttö puolestaan poistaa tarpeen muistaa muuttujien sekä konekäskyjen osoitteita.

Kuva 1 selventää symbolisten konekielten määrittelymien symbolien sekä ohjelmoijan määrittelymien tunnisteyden eron. Esimerkkikoodi on TTK91 (Ttk, 1991) virtuaaliprosessorille tehty ohjelma, joka tulostaa käyttäjälle luvut 0...5. Keltaisella pohjalla olevat symbolit ovat niin sanottuja tunnisteytiä, joilla voi olla eri arvo käännoškerrasta ja kääntäjästä riippuen. Kaikki harmaalla pohjalla oleva koodi käännetään siis täysin ennalta määrätysti.

	LOAD	R1,	=	0
	STORE	R1,		X
LOOP	LOAD	R1,	=	5
	COMP	R1,		X
	JGRE			END
	LOAD	R1,		X
	OUT	R1,	=	CRT
	ADD	R1,	=	1
	STORE	R1,		X
	JUMP			LOOP
END	NOP			

Kuva 1: TTK91 esimerkkikoodi

3 Historiaa korkean tason kielten kääntäjistä

Korkean tason ohjelmointikielellä tarkoitetaan tässä dokumentissa ohjelmointikieltä, jossa lähdekielikieli sekä siitä käännettävä konekieli ovat selkeästi eri abstraktiotasoilla ja kääntäminen edellyttää muutakin, kuin mekaanista sanojen vaihtamista ennalta määrättyjen sääntöjen perusteella. Tämän määritelmän perusteella korkean tason ohjelmointikielillä tarkoitetaan tässä dokumentissa niitä kieliä, jotka eivät ole symbolisia konekieliä.

3.1 Täsmällisten kuvausjärjestelmien kehitys

Tietojenkäsittelytieteilijät ovat jo tietokoneiden alkua ajoista lähtien yrittäneet kuvailla ohjelmien suoritusta ja algoritmeja konekieltä abstraktimmalla tasolla. Alan Turingin julkaisussa vuonna 1936 esitettiin määritelmä tietojenkäsittelijöiden hyvin tuntemasta laskentalaitteesta, Turingin koneesta. Laitteen yhteydessä määriteltiin myös matemaattinen esitystapa, jolla sen toimintaa voitiin täsmällisesti kuvailla. Vaikka esitystapa oli vaikea eikä kyseisiä Turingin esittelemää laitetta ollut kuin teoriassa, Turingin esitystapa edusti kehittyneintä formaalia kuvausta, 'kieltä', joka siihen aikaan oli olemassa.

Toisen maailmansodan jälkeen vuonna 1945 saksalainen Konrad Zuse aloitti oman tietokoneohjelmien kuvailuun tarkoitetun kielen Plankalkülin kehittämisen. Zusen sanoin Plankalkülin tarkoitus oli luoda puhtaasti formaali esitystapa mille tahansa laskentaongelmalle (Knuth and Pardo, 1976: 10). Tässä hän onnistuikin varsin hyvin. Plankalkulissa voidaan määritellä aritmetiikan ja ohjausrakenteiden lisäksi rajaton määrä sisäkkäisiä tietorakenteita ja Zusen työhön viitataan usein ensimmäisenä korkean tason ohjelmointikielenä. Vaikka kyseessä oli huomattavan edistyksellinen järjestelmä, se ei kuitenkaan vaikuttanut ohjelmointikielten kehitykseen juuri lainkaan. Zusen artikkelit julkaistiin vasta vuonna 1972 muiden, kehittyneempien kielten jo olemassa ollessa. Vaikka Plankalkülille toteutettiin kääntäjä, ei sitä juuri koskaan käytetty koska silloin oli jo Plankalküliä huomattavasti kehittyneempiä ohjelmointikieliä.

Samoihin aikoihin Zusen kanssa Yhdysvaltalaiset Herman Goldstine ja John von Neumann koittivat ratkaista samaa ongelmaa. Heidän ratkaisunsa algoritmien ja tietokoneohjelmien kuvaamiseen oli varsin erilainen. Von Neumann ja Goldstine esittivät ratkaisuksi lohkokaaaviota (flow diagram), esitystapaa jossa ohjelmat kuvataan nuolien ja laatikoiden avulla.

Vuonna 1946 Marylandissa työskennellyt Haskell B. Curry kehitti ENIAC tietokoneelle aikaansa nähden monimutkaista ohjelmaa. Curryn työ ENIACIN parissa sai hänet ehdottamaan formalismia ohjelmistojen toiminnalle. Hänen formalisminsa perustui uuteen ajatukseen ohjelman suorituksen lohkomaisesta rakenteesta, mitä hän nimitti divisiooniksi. Divisioonien tulisi olla rakennettu niin että niiden laskenta olisi toisistaan riippumatonta. Tämän voisikin rinnastaa esimerkiksi C-kielen paikallisiin tietorakenteisiin ja käännösyksiköihin perustuvaan suoritukseen. Curryn formalismi oli kuitenkin hieman luonnoton sillä suoritusyksiköillä oli useita lopetuskohтия sekä nykykielistä poiketen useita aloituskohтия. Historiallisesti työ oli kuitenkin merkittävä, sillä se sisälsi algoritmeja joilla kuvauksesta pystyttiin tuottamaan konekoodia. Näitä rekursiivisia — vaikkakin toteuttamatta jääneitä — algoritmeja voidaankin pitää ensimmäisinä koodin generointiin tarkoitettuina algoritmeina.

3.2 Kohti ensimmäisiä kääntäjiä

Millekään aiemmin mainituista ohjelmointikielistä ei tähän mennessä oltu toteutettu kääntäjiä. Ne toimivat ohjelmoijien käsitteellisenä apuna auttaen ohjelmien suunnittelussa, mutta jättäen toteutuksen ihmisille. Tästä huolimatta ne kaikki olivat merkittäviä askeleita kohti parempia ohjelmointikieliä sekä niiden kääntäjiä. Ilman täsmällisiä esitystapoja ei koodin generointi ikinä olisi tullut mahdolliseksi.

Ensimmäinen korkean tason ohjelmointikieli, jolle toteutettiin tulkki oli Short Code. Sitä kehitti John W. Mauchly vuonna 1949 ja William F. Schmitt toteutti sille tulkin (Knuth and Pardo, 1976: 23). Tulkki toimi alkuun BINAC tietokoneella mutta se ohjelmoitiin myöhemmin myös UNIVACille. Yksityiskohtia Short Coden toiminnasta ei ikinä julkaistu, joten sen tarkemmasta toiminnasta ei ole tietoa. Vuonna 1955 julkaistusta ohjelmoijille tarkoitetussa manuaalissa kerrotaan kuitenkin kuinka ohjelmaa voidaan käyttää. Short Code oli siis algebrallinen tulkki, joka osasi suorittaa aritmeettisia laskutoimituksia ilman konekielistä ohjelmointia. Ohjelma luki syötettä ja suoritti vastaavat toiminnot ajatulla laitteistolla.

1950-luvun alussa Heiniz Rutishauser ja Corrado Böhm työskentelivät Zürichin teknillisessä yliopistossa Sveitsissä. Vaikka he työskentelivät samassa paikassa ja saman aiheen parissa, eivät he työskennelleet yhdessä. Rutishauser julkaisi 1952 artikkelin, jossa hän kuvasi hypoteettisen tietokoneen sekä siinä toimivan kääntäjän kehittämälleen ohjelmointikielelle. Julkaisu oli merkittävä, sillä siinä kuvattiin ensimmäistä kertaa menetelmä kääntäjien toteuttamisesta sekä koodin generoinnista.

Rutishauserin kollega Corrado Böhm kehitti myös ohjelmointikieltä sekä tämän kääntäjää. Hänen julkaisunsa oli Rutishauserin julkaisua vieläkin merkittävämpi, sillä hän oli toteuttanut kääntäjän tämän omalla kielellä. Böhmin kieli ei kuitenkaan osannut käsitellä muita kuin positiivisia kokonaislukuja, joten sen käyttöarvo jäi melko pieneksi. Kääntäjien teorian kehityksen kannalta se oli kuitenkin korvaamaton. Böhmin kääntäjä kykeni tarkistamaan koodin syntaksia lineaarisessa ajassa kun Rutishauserin kääntäjä toimi suuruusluokassa n^2 . Lisäksi Böhmin kääntäjä halitsi matemaattisten operaattoreiden sidontajärjestyksen, sekä osasi käsitellä sulkeita aritmeettisissa lausekkeissa. Lisäksi Böhm oli ensimmäinen tietojenkäsittelijä, joka todisti matemaattisesti ohjelmointikielensä voivan laskea minkä tahansa laskettavan funktion.

Vaikka Rutishauser ja Böhm olivat kumpikin valmistaneet omat kääntäjänsä, pidetään ensimmäisenä 'oikeana' kääntäjänä silti Alick E. Glennien 1952 valmistamaa AUTOCODE ohjelmistoa. Aiemmista kääntäjistä poiketen AUTOCODE toteutettiin oikealle laitteistolla ja sen tuottama konekieli oli oikeasti suoritettavissa. AUTOCODEa pystyttiin siis käyttämään oikeiden, käyttökelpoisten ohjelmien tekemiseen (Knuth and Pardo, 1976: 42).

Vuoden 1954 alussa John Backus rupesi kehittämään kokoamansa kehittäjätiimin kanssa automaattisen ohjelmoinnin järjestelmää. Järjestelmän oli tarkoitus olla hyvin kehittynyt, joten suureksi haasteeksi muodostui järjestelmän saaminen tarpeeksi tehokkaaksi. Loppuvuodesta 1954 kehittäjäryhmä julkaisi suunnitelman järjestelmästä 'The IBM Mathematical FORMula TRANstating system' — FORTRAN. Kuten jo aiemmin oli todettu, tehokkaan koodin tuottaminen ei ollut lainkaan helppoa. Ryhmän julkaisu alkoikin painottamalla sitä tosiasiaa, että FORTRAN oli tehokas. Aiemmin ohjelmoijien tuli valita helpon ohjelmoinnin ja hitaan suorituksen tai työlään ohjelmoinnin ja nopean suorituksen väliltä, mutta FORTRANin tarjoaisi parhaat puolet molemmista (IBM, 1954: 1). FORTRAN 0 dokumentti esittää myös ensimmäisen yrityksen esittää ohjelmointikielen syntaksi täsmällisesti. Tätä voidaan pitää Backuksen myöhemmin esittelemän kielioppimuodon Backus Naur Formin (BNF) edeltäjänä.

Kun FORTRAN kaksi ja puoli vuotta myöhemmin saatiin toteutettua, oli se aikansa tehokkain sekä monipuolisin ohjelmointikieli. FORTRAN tuotti kohtuullisen tehokasta koodia ja kehittäjät sanoivat sen olevan lähes yhtä tehokasta kuin hyvän ohjelmoijan kirjoittama symbolinen konekieli. FORTRANissa oli myös paljon ominaisuuksia, joita ei oltu aiemmin nähty. Se oli esimerkiksi ensimmäinen ohjel-

mointikieli, jossa muuttujien nimet voivat olla useamman merkin pituisia (Knuth and Pardo, 1976: 62).

Ensimmäisen julkaisun jälkeen FORTRANissa oli kuitenkin useita ongelmia. Virheitä oli paljon ja eräs FORTRANIN kehittäjästä, Saul Rosen, sanoikin ettei uskonut FORTRANin ikinä tulevan toimimaan (Rosen, 1964: 4????). Vaikeuksista huolimatta FORTRANista tuli hyvin suosittu ja sitä käytettiin enemmän kuin oltiin osattu odottaa.

Taulukko 1 tiivistää aiemmissa kappaleissa esiteltyjen ohjelmointikielten merkittävimmät piirteet. Lisäksi taulukossa esitellään ohjelmointikielten nimet sekä päätekijät.

Taulukko 1: Yhteenveto ohjelmointikielten ja kuvausjärjestelmien kehityksestä

Kieli	Kehittäjä	Ensimmäinen
Plankalkül	Zuse	Ohjelmointikieli, Hierarkkinen data
Virtauskaaviot	Goldstine, Von Neumann	Hyväksytty ohjelmointimetodologia
Short Code	Mauchly	Toteutettu korkean tason ohjelmointikieli
Formules	Böhm	Samalla kielellä kirjoitettu kääntäjä
AUTOCODE	Glennie	Käyttökelpoinen kääntäjä
FORTRAN I	Backus	I/O formaatti, kommentit, globaali optimointi

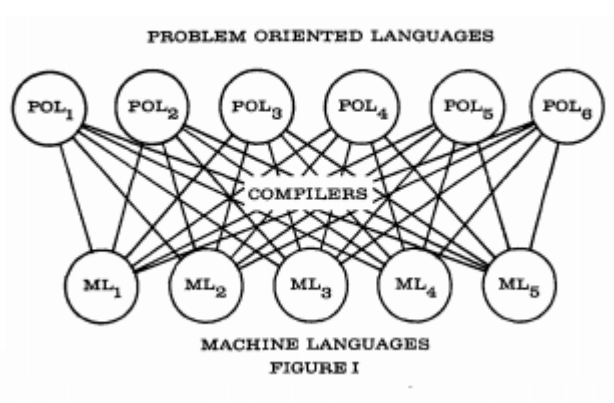
4 Välikielet ja ristiinkääntäminen

Jo 1950-luvun lopulla sekä käytettävissä olevien ohjelmointikielten että erillaisten laitteistojen määrä oli kasvanut merkittävästi. Korkean tason ohjelmointikielillä ohjelmoinnista oli tullut yleistä ja sen hyödyt olivat laajalti tunnettu (Strong et al., 1958: 12). Uusien laitteistojen kehitys oli nopeaa ja tyypillisesti laitteistoja uusittiin viimeistään viiden vuoden välein (ibid). Vaikka korkean tason ohjelmointikielten käyttö vähensi tarvetta sovellusten uudelleenohjelmoinnille, muodostui kääntäjien kehitys merkittäväksi ongelmaksi. Tehokkaan kääntäjän kehittämiseen kuluva aika oli samaa suuruusluokkaa laitteistojen käyttöiän kanssa, joten hyvien kääntäjien valmistuttua oli laitteisto jo vanhentunut (ibid). Ongelmaa koitettiin välttää erilaisilla käännöstekniikoilla. Yleiseksi tavaksi muodostui ristiinkääntäminen uusille laitteistoille.

4.1 Ristiinkääntäminen

Tavallisesti tietokonejärjestelmissä käytettävät kääntäjät tuottavat konekieltä samalle laitteistolle, jossa niitä ajetaan. Ristiinkääntämisellä (cross compiling) tarkoitetaan sitä että, kohdelaitteisto on jokin muu, kuin käännöstä suorittava laitteisto (GNU, 2016). Ristiinkääntämisen mahdollistamiseksi tehdään, vanhalla laitteistolla jo olemassa olevia ohjelmointikieliä käyttäen kääntäjiä, jotka tuottavat ohjelmia uusille tietokoneille. Saaduilla ristiinkääntäjillä pystytään ohjelmoimaan kääntäjiä jotka toimivat uudessa laitteistossa.

Ongelmana edellä mainitussa menetelmässä on suuri, jatkuvasti kasvava joukko sekä lähde- että kohdekieliä. Jo yhden uuden kääntäjän toteuttaminen on varsin työlästä joten uuden kääntäjän tekeminen kaikille halutuille laitteistoille vaatisi massiivisesti aikaa ja resursseja. Oletetaan että korkean tason ohjelmointikielten määrä on N ja laitteistojen määrä on M . Tällöin tarvittavien kääntäjien määrä on $N \times M$ ja se kasvaa neliöllisesti kohde- tai lähdekielten kasvaessa. Kuva X esittää tilannetta jossa on kääntäjät kuudelle eri ohjelmointikielelle viiteen eri laitteistoon.



Kuva 2: Ristiinkääntäminen usealta lähdekieleltä usealle kohdekielelle. Lähde: (Jr., 1961: 378).

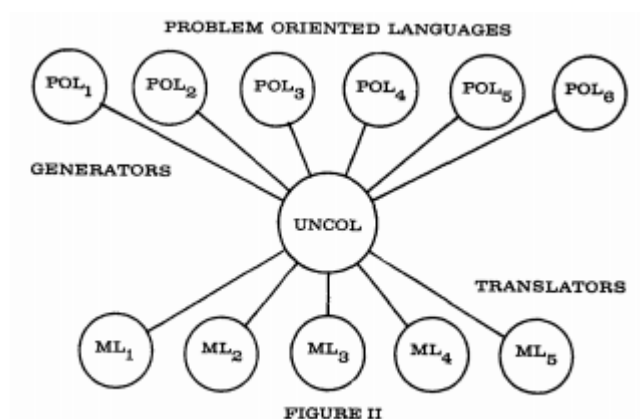
4.2 Välikielet

Mikäli olisi olemassa sellainen ohjelmointikieli, jolle olisi toteutettu kääntäjä kaikille mahdollisille laitteistoille, voitaisiin korkean tason ohjelmointikielet kääntää mille

tahansa laitteistolle tätä kieltä apuna käyttäen. Tällaisella menetelmällä tarvittavien kääntäjien määrä vähenisi huomattavasti. Uudet ohjelmointikielet saataisiin toimimaan millä tahansa laitteistolla vain yhen uuden kääntäjän avulla. Uuden ohjelmointikielen kääntäjä kääntäisi sitä kyseiselle välikielelle, josta se voitaisiin jo olemassa olevien kääntäjien avulla kääntää mille tahansa laitteistolle.

Idea tunnettiin alunperin nimellä UNCOL (UNiversal Computer Oriented Language) (Strong et al., 1958: 14). UNCOL ei ollut mikään suunniteltu ohjelmointikieli, vaan pikemminkin idea välikielestä, jonka avulla kääntäjiä voisi toteuttaa pienemmällä vaivalla. Siitä puhuttiin jo vuonna 1954, eikä konseptin alkuperäistä keksijää tunneta (ibid).

Mikäli jonkinlainen UNCOL — kaikkien kääntäjien tuntema välikieli — olisi olemassa, vähenisi ristiinkääntämiseen tarvittavien kääntäjien määrä huomattavasti. Jos ohjelmointikielten määrä on N ja laitteistojen määrä M , tarvittaisiin tällaisessa välikieliratkaisussa vain $N+M$ uutta kääntäjää (Speetjens, 1976: 15). Toisin sanoen kääntäjien määrä kasvaisi lineaarisesti ohjelmointikielten ja laitteistojen suhteen. Kuvassa X esitetään tilanne välikielen avulla.



Kuva 3: Ristiinkääntäminen välikielillä. Lähde: (Jr., 1961: 378).

UNCOL:ksi on ehdotettu useita eri välikieliä, mutta yhdestäkään ei ole tullut niin käytettyä, että sitä voitaisiin sanoa universaaliksi. Eräs UNCOL:ksi ehdotettu kieli oli Melvin Conwayn 1958 julkaisema välikieli. Conwayn julkaisussa "Proposal for an UNCOL" määritellään matalan tason välikieli, jota hän kutsuu SML:ksi (Simple Machine Language) (Conway, 1958: 5). Myöhemmin

Kaikilla laitteilla toimivan välikielen tulisi olla täysin laitteistoriippumaton. Laitteistoriippumattomalla kielellä tarkoitetaan, että se voidaan kääntää tehokkaasti mille tahansa laitteistolle (Brown, 1972: 1060). Koska erilaisia laitteistoja on hyvin suuri määrä, ei mikään kieli ole täysin laitteistoriippumaton (ibid). Näin ollen kaikille laitteistoille käännettävän välikielen toteutus on melko epärealistista. Useat laitteistot ovat kuitenkin tarpeeksi samankaltaisia, tehokkaan välikielen kehittämiseksi, eikä joukosta puutu kovin moni oikeassa käytössä oleva tietokone (ibid).

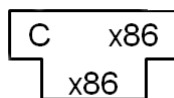
Vaikka välikielet ovat usein laitteistoläheisiä ja ne käännetään suoraan jollekin konekielelle, voisi välikieli teoriassa olla hyvinkin korkeatasoinen. Jos jollekin laitteistolle on olemassa hyvä jonkin korkean tason ohjelmointikielen kääntäjä, voisi välikieli olla myös korkean tason ohjelmointikieli, jonka kohdekielenä ei olisi mikään konekieli vaan kyseinen toinen korkean tason ohjelmointikieli (Brown, 1972: 1060). Koska tällainen tekniikka tuottaa käytännössä aina huonotasoista koodia, käännetään välikielet lähes poikkeuksetta, vaikka olisivatkin korkean tasoisia ohjelmointikieliä, suoraan konekieliksi (ibid).

5 T-kaaviot

Kääntäjien suunnittelussa ja mallintamisessa on käytetty useita erilaisia kaavioita. Käytetyin ja tunnetuin lienee Harvey Bratmanin 1961 ehdottama T-kaavio, jonka hän kehitti UNCOL-kaavion korvaajaksi (Bratman, 1961).

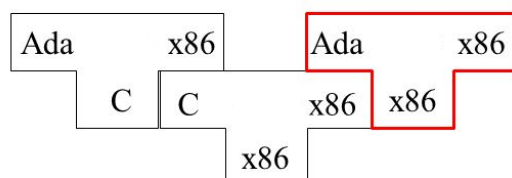
5.1 Alkuperäinen Bratman-kaavio

Harvey Bratman esittelemä kaavio kuvaa yksittäistä kääntäjää. Siitä käy ilmi kääntäjän ymmärtämä kohde- ja lähdekieli sekä kieli, jolla kääntäjä toimii. Kaaviota kutsutaan Bratman-kaavioksi tai T-kaavioksi. Jälkimmäinen nimi tulee kaavion muodosta, jossa T-kirjaimen muotoisessa alueessa vasempaan pätyyn merkitään lähdekieli, oikeaan pätyyn kohdekieli ja alaosaan kieli jolla kääntäjä toimii.



Kuva 4: T-kaavio, joka kuvaa x86-arkkitehtuurilla toimivaa C-kääntäjää, jonka kohdekieli on x86-konekieli.

Kaavioita toisiinsa liittämällä voidaan havainnollistaa monimutkaisiakin kääntäjillä suoritettavia toimintaketjuja. Kuvassa 3 oletetaan että, käytössä on C-kielinen kääntäjä, joka kääntää Ada-kieltä x86-konekielelle. Lisäksi käytössä on edellisen kuvan esimerkissä oleva x86-arkkitehtuurilla toimiva C-kääntäjä, jonka kohdekieli on x86-konekieli. Näiden kahden kääntäjän avulla voidaan tuottaa x86-alustalla suoritettava Ada kääntäjä, jonka kohdekieli on x86. Kahden ensimmäisen kääntäjän yhteistyöllä saadaan siis kolmas kääntäjä. Huomattavaa on, että kaavion alin kääntäjä toimii aina jossakin todellisessa laitteistossa, eikä täten voi olla muu kuin jonkin laitteiston ymmärtämä konekieli.

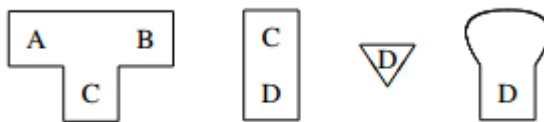


Kuva 5: C-kielisen Ada-kääntäjän ja x86:lla toimivan C kääntäjän avulla voidaan tuottaa x86:lla toimiva ada-kääntäjä

5.2 Earleyn ja Sturginsin merkintätapa

T-kaavio on melko yksinkertainen, joten siitä on kehitetty paranneltuja vaihtoehtoja. Jay Earley ja Howard Sturgis laajensivat Bratmanin kaaviota lisäämällä siihen ominaisuuksia (Earley and Sturgis, 1970: 607-616). Alkuperäisen T-kaavion kuvatessa

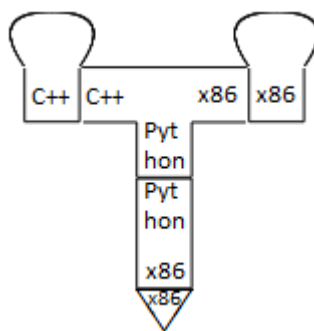
vain kääntäjiä, Earleyn ja Sturgisin kaavioissa pystyi kuvaamaan myös tulkkeja. Lisäksi kaavioon sisällytettiin suoritusta kuvaava osa sekä sovellusohjelmaa kuvaava kaavio. Heidän kaavioissaan määritellään kuvan 4 mukaiset elementit.



Kuva 6: Earleyn ja Sturgisin T-kaavioelementit. Kuvan lähde (Mogensen, 2010: 282).

Kuvan vasemman puoleisin elementti vastaa Earleyn ja Sturgisin merkintätavassa alkuperäistä T-kaaviota. Se kuvaa C-kielistä kääntäjää joka kääntää kieleltä A kielelle B. Suorakaiteen muotoinen kahdesta osasta koostuva kaavio kuvaa tulkkeja, joka tulkkaa kieltä C ja toimii kielellä D. Jotta käännös voitaisiin suorittaa, tulee pohjimmaisena kielen olla suoritettavissa jollakin oikealla laitteistolla. Tätä suoritusta kuvaamaan lisättiin yksiosainen kolmiomerkinä. Kolmion sisällä lukee mitä konekieltä suoritus ymmärtää. Esimerkkikuvassa konekieli on D. Lampun muotoinen kaavio tarkoittaa jotain sovellusohjelmaa tai määrittelemätöntä laskentaa. Kaavion sisällä lukee ohjelmointikieli, jolla sovellus on ohjelmoitu tai jolle se on käännetty. Esimerkkikuvassa kieli on D (Mogensen, 2010: 282).

Kyseisiä kaavioita voi yhdistellä hyvinkin monimutkaisiksi rakenteiksi. Kuvan 5 esimerkissä on kaavio, joka kuvaa C++-kielisen ohjelman kääntämistä x86-konekielelle. Suoritus tapahtuu x86-laitteistolla ja käännöksen tekee pythonilla toimiva kääntäjä, joka tulkkataan x86-laitteistossa. Huomioitava T-kaavioiden käytössä on, että kaavioiden vierekkäisten osien kielten tulee täsmätä toisiinsa. Kuvan esimerkin käännös ei voi ottaa lähdeohjelmakseen muuta kuin C++:llä kirjoitetun ohjelman.



Kuva 7: Esimerkki Earleyn ja Sturgisin kaavioiden yhdistelemisestä.

6 Kääntäjien rakenne ja ketjutus

Koska kääntäjien ohjelmointi matalan tason ohjelmointikielillä on erittäin vaivalloista (Mogensen, 2010: 281), suositetaan niiden kehittämisessä korkean tason ohjelmointikielten käyttöä. Yksi vaihtoehto on ohjelmoida laitteistolle kääntäjä jollakin kyseiselle laitteistolle jo olemassa olevalla ohjelmointikielellä. Tämä ei kuitenkaan aina ole mahdollista. Esimerkiksi, jos kyseessä on uusi prosessoriarkkitehtuuri, eikä tälle vielä ole kääntäjiä, on ohjelmointi tehtävä jollakin muulla tavalla. Kääntäjän joka sekä kääntää että on suoritettavissa kyseisellä laitteistolla tuottamiseksi, eräs yleisesti käytetty menetelmä on ketjutus (bootstrapping) (Mogensen, 2010: 281).

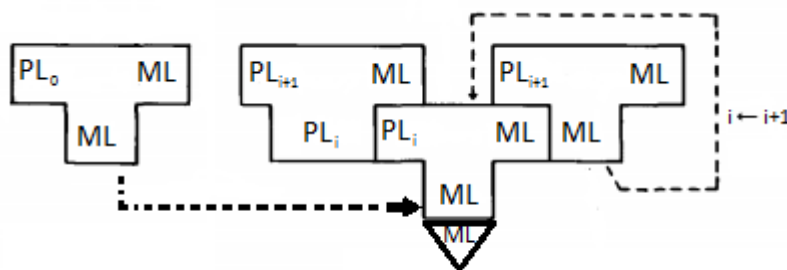
Ketjutus on tekniikka, jonka ydinajatus on kääntää jokin kääntäjä sillä itsellään (Mogensen, 2010: 281) ja tavanomaisen kääntäjän sanotaan olevan ketjutettu jos se kääntää itsensä (Appel, 1994).

6.1 Iteratiivinen ketjutus

Perinteinen ongelma kääntäjien kehityksessä on tilanne, jossa kääntäjä pitää ohjelmoida ilman muita työkaluja ja kääntäjiä. Iteratiivisessa ketjutus- menetelmässä kääntäjä kehitetään kahdessa osassa. Ensin kääntäjästä tehdään hyvin suppea versio, joka osaa kääntää tavoitellusta ohjelmointikielestä vain pienen osajoukon (Mogensen, 2010: 287). Tämä ensimmäisen vaiheen kääntäjä voidaan ohjelmoida millä tahansa ohjelmointikielellä, mutta mikäli muiden ohjelmointikielten kääntäjiä ei ole saatavilla, se joudutaan tekemään konekielellä. Toisessa osassa kääntäjä ohjelmoi-

daan sen itsensä ymmärtämällä kielellä, eikä muita ohjelmointikieliä enää tarvita. Tällöin jo olemassa oleva kääntäjä voi kääntää seuraavan version itsestään. Koska kieli on tässä vaiheessa vielä hyvin vajavainen, olisi valmiin ohjelmointikielen toteuttaminen heti ensimmäisen vaiheen jälkeen erittäin työlästä tai mahdotonta. Toista vaihetta suoritetaan useita kertoja siten, että kääntäjän tuntemaa ohjelmointikieltä kasvatetaan ja siitä käännetään uusi kääntäjä, jonka avulla kieltä taas kasvatetaan, ja niin edelleen. Huomioitavaa prosessissa on se, että kääntäjän uusi versio on kehitettävä aina käyttäen vanhan version tuntemaa ohjelmointikieltä.

Kuva 5 esittää edellä mainittua tilannetta. Ensimmäisessä vaiheessa kääntäjän ensimmäinen versio ohjelmoidaan konekielellä (ML, "Machine Language") ja tämän jälkeen ohjelmointikieltä (PL, "Programming Language") laajennetaan iteratiivisesti, kunnes tavoiteltu kieli on saavutettu.



Kuva 8: Iteratiivinen bootsträppäys ilman muita ohjelmointikieliä.

Eräs ongelma edellä kuvatussa prosessissa on kielen kehitykseen soveltuvan sopivan osajoukon löytäminen (Reynolds, 2003: 176). Mikäli ohjelmointikieltä laajennetaan vain vähän, on päivitys lähes hyödytön. Liian monimutkaisen kielen toteuttaminen yksinkertaisella osajoukolla on puolestaan erittäin työlästä.

6.2 Ketjutus olemassa olevien kääntäjien avulla

Ketjutus-menetelmää käytetään usein myös yhdessä ristiinkääntämisen (cross compiling) kanssa (Reynolds, 2003: 175). Tällöin tavoitteena on siirtää jo olemassa oleva kääntäjä toimimaan toisessa laitteistossa. Oletetaan, että laitteistolle, joka suorittaa konekieltä ML1 on toteutettu kielen PL kääntäjä. Haluttaessa kielen PL kääntäjä

6.3 Ketjuttamista helpottava kääntäjä rakenne

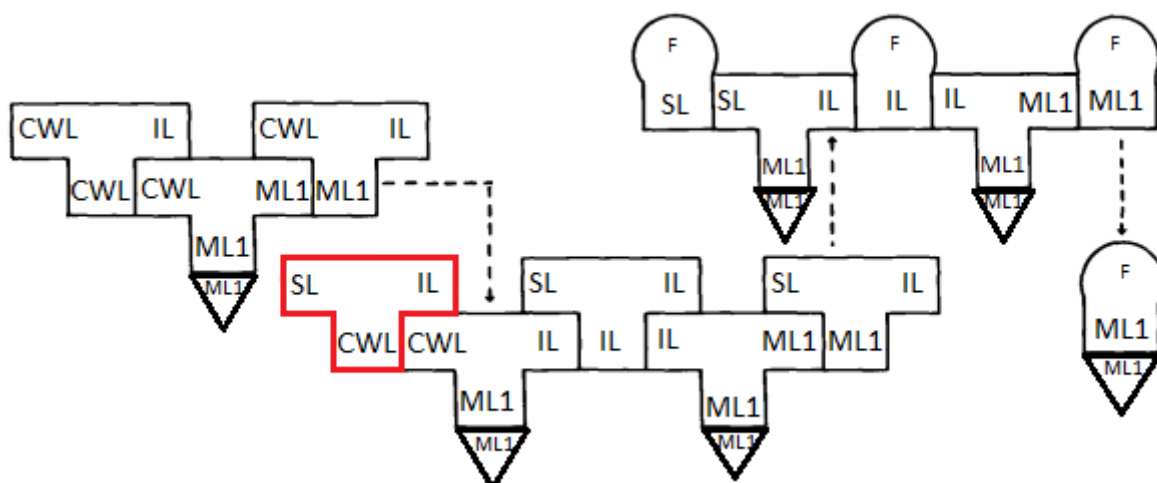
Vielä 1970-luvulla kääntäjät kirjoitettiin hyvin tarkasti tiettyä lähdekieltä ja laitteisto varten (Guilan et al., 2002: 38). 1980-luvulla ruvettiin kehittämään enenevässä määrin kääntäjäkokonaisuuksia, jotka kykenevät kääntämään useita eri lähdekieliä (ibid). Tällaisissa systeemeissä kääntäjän sisäinen rakenne oli jaettu lähdekielestä riippuvaiseen etuosaan (frontend) sekä näiden jakamaan takaosaan (backend). Tämän kaltainen rakenne vähensi tarvetta kirjoittaa uutta koodia ja näin ollen pienensi kääntäjien kehityksestä koituvia kustannuksia. 80-luvun lopulla oli jo hyväksytty ajatus siitä, että kääntäjien kehitys usealle ohjelmointikielelle ja/tai laitteistolle on tehokkaampaa sekä kilpailukykyisempää (ibid).

Eräs tyypillinen tapa kääntäjien ohjelmoinnissa onkin jakaa sen sisäistä rakennetta sekä suorittaa käännös useassa eri vaiheessa. Apuna käytetään usein jotakin niinsanottua välikieltä (intermediate language) (Speetjens, 1976: 15). Näin haluttu lähdekieli voidaan kääntää helpommin eri laitteistoille.

Olettaen, että jollekin välikielelle IL (Intermediate Language) ja jollekin korkean tason ohjelmointikielelle on olemassa tietyllä laitteistolla toimivat kääntäjät, voidaan halutun lähdekielen käännös suorittaa kaksivaiheisesti toteuttamalla korkean tason ohjelmointikielellä kääntäjä lähdekieleltä välikielelle (Earley and Sturgis, 1970: 610). Uuden lähdekielen toteutus sisältää siis vain korkean tason ohjelmointikielellä tehdyn kääntäjän käytetylle välikielelle, eikä koodin generointivaihetta tarvitse kirjoittaa uudestaan.

Kuva 7 demonstroi usein käytettyä rakennetta (Earley and Sturgis, 1970: 610), jolla kääntäjän jatkokehitystä voidaan helpottaa huomattavasti.

Uuden ohjelmointikielen kääntäjän toteuttaminen kyseisen kääntäjän avulla on mahdollista vain yhden komponentin (punainen) uudelleenohjelmoinnilla. Vaihtamalla toisen vaiheen ensimmäinen kääntäjä kääntämään uudelta lähdekieleltä (SL, "Source Language") välikielelle, vaihtuu myös seuraavissa komponenteissa olevat lähdekielet uuteen kieleen. Uusi kääntäjä tulee toteuttaa samalla, ohjelmointikielellä, jolla aiempi komponentti oli toteutettu (CWL, "Compiler Writing Language). Esimerkiksi haluttaessa kyseisellä järjestelmällä toteuttaa konekielellä ML1 suoritettava C-kääntäjä, tulee ohjelmoida vain sellainen osa, joka on kirjoitettu kielellä CWL ja kääntää C-kieltä välikielellä IL. Sama pätee muihinkin toteutettaviin ohjelmointikieliin.



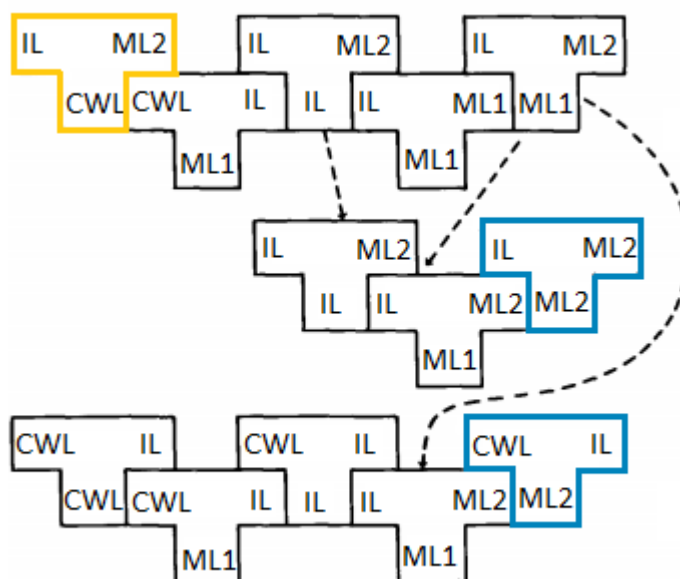
Kuva 10: Kääntäjäkokonaisuus.

Pelkkää välikieltä apuna käyttäen uuden ohjelmointikielen toteutus ei vaatisi näin monimutkaista rakennetta. Kyseinen rakenne mahdollistaa kuitenkin sekä tehokkaan ristiinkääntämisen, että itsensä ketjuttamisen uudelle laitteistolle. Jos kääntäjäkokonaisuuden halutaan kääntävän ymmärtämäänsä lähdekieltä uudelle laitteistolle, jonka konekieli on ML2, joudutaan siihen edelleen kirjoittamaan vain yksi uusi komponentti. Tämä osa kirjoitetaan kielellä CWL ja se kääntää välikieltä IL uuden laitteiston ymmärtämälle konekielelle ML2. Kun kyseinen osa käännetään jo olemassa olevalla kääntäjällä, saadaan tulokseksi samassa laitteistossa toimiva kääntäjä välikieltä uudelle konekielelle. Saadulla ristiinkääntäjällä on siis mahdollista kääntää lähdekieltä konekielelle ML2, mutta sen suoritus tapahtuu edelleen kielen alkupe- räistä konekieltä ML1 ymmärtävässä laitteistossa.

Ristiinkääntämisen mahdollistavan uuden osan avulla myös koko kääntäjärakenteen ketjuttamisen uuteen laitteistoon on mahdollista ilman lisäohjelmointia (Earley and Sturgis, 1970: 610). Kuva 8 havainnollistaa tätä prosessia. Ensimmäisessä vaiheessa jo olemassaolevien komponenttien sekä uuden CWL:llä kirjoitetun välikielikääntäjän (keltainen) avulla tuotetaan seuraavissa vaiheissa tarvittavat komponentit.

Näiden komponenttien avulla voidaan toisessa vaiheessa tuottaa ML2-konekielellä toimiva ja siihen kääntävä välikielikääntäjä. Kolmannessa vaiheessa tuotetaan ensimmäisessä vaiheessa saadun välikielikääntäjän avulla ML2-kielellä toimiva CWL-

kääntäjä, joka tuottaa välikieltä.



Kuva 11: Käänätäkokonaisuuden ketjutus uudelle laitteistolle.

Toisessa ja kolmannessa vaiheessa saadut kääntäjät (sininen) sijoittamalla alkuperäiseen ohjelmistoon, saadaan sekä ML2-kielillä toimivat, että sille kääntävät komponentit. Ainoa komponentti jota ei saada suoraan sijoittamalla on ensimmäisen vaiheen toinen komponentti. Uudella konekielellä toimivat kääntäjät $CWL \rightarrow IL$ ja $IL \rightarrow ML2$ vastaavat kuitenkin tätä komponenttia.

Käsitelty kääntäjäkokonaisuus on monimutkainen, mutta myös tehokas. Sekä uuden lähdekielen toteutus alkuperäisellä laitteistolla, että koko kääntäjäohjelmiston siirtäminen uuteen laitteistoon voidaan saavuttaa vain pienen osan uudelleenohjelmoinnilla. Lisäksi uudelleenohjelmointi voidaan suorittaa jo aiemmin käytetyllä korkean tason ohjelmointikielellä (Earley and Sturgis, 1970: 610).

7 Yhteenveto

Tietojenkäsittelytieteessä kääntäjä tarkoittaa ohjelmaa joka kääntää lähdekielisen ohjelmakoodin kohdekieliseksi ohjelmakoodiksi. Koska ennen muun kuin konekielten kehitystä ohjelmointi tapahtui suoraan laitearkkitehtuurin ymmärtämällä muo-

dolla, ei tarvetta kääntäjille ollut. Koska konekoodin ohjelmointi oli varsin työlästä, kehitettiin avuksi symbolisia konekieliä, joissa tietyt binäärijonot oli korvattu paremmin muistettavilla tekstuaalisilla symboleilla.

Korkeamman tason ohjelmointikielet kehittyivät ohjelmoijien tarpeesta kuvata ohjelmistojen toimintaa korkeammilla abstraktiotasoilla. Täsmälliset kuvausjärjestelmät kehitettiin alunperin ilman ajatusta kääntäjistä taikka automaattisesta koodin generoinnista. Vaikka kuvausjärjestelmien sekä koodia generoivien algoritmien kehitys oli alkuun vain teoreettista tutkimista, huomattiin potentiaali niiden tehokkaaseen käyttöön varsin pian. Tämän jälkeen ohjelmointikieliä ruvettiin kehittämään varta vasten automaattisen koodin generoinnin takia ja ensimmäiset oikeasti hyödylliset kielet sekä niiden kääntäjät kehitettiin.

Lähteet

- [1] (1991). Ttk91 reference. https://www.cs.helsinki.fi/group/titokone/ttk91_ref_en.html. Accessed: 2016-10-21.
- [2] (2016). Cross-compilation. https://www.gnu.org/savannah-checkouts/gnu/automake/manual/html_node/Cross_002dCompilation.html. Accessed: 2016-12-02.
- [3] Appel, A. (1994). Axiomatic bootstrapping: A guide for compiler hackers.
- [4] Bauer, F. L. (1974). *Compiler Construction – An Advanced Course*. Springer-Verlag, Berlin Heidelberg GmbH.
- [5] Bratman, H. (1961). An alternative form of the 'uncol' diagram.
- [6] Brown, P. J. (1972). Levels of language for portable software.
- [7] Conway, M. (1958). Proposal for an uncol.
- [8] Earley, J. and Sturgis, H. (1970). A formalism for translator interactions.
- [9] Goldstine, H. and Von Neumann, J. (1947). Planning and coding problems for an electronic computing instrument.
- [10] Guilan, D., Jinlan, T., Suqin, Z., Weidu, J., and Jun, D. (2002). Retargetable cross compilation techniques – comparison and analysis of gcc and zephyr.
- [11] IBM (1954). Specifications for the ibm mathematical formula translating system, fortran Programming Research Group, I.B.M Applied Science Div.
- [12] Jr., T. B. S. (1961). A first version of uncol.
- [13] Knuth, D. and Pardo, L. (1976). *The early development of programming languages*. Stanford University.
- [14] Mogensen, T. (2010). *Basics of Compiler Design*. lulu.
- [15] Reynolds, J. (2003). Bootstrapping a self compiling compiler from machine x to machine y.
- [16] Rosen, S. (1964). Programming systems and languages, a historical survey.
- [17] Salomon, D. (1993). *Assemblers and Loaders*. Ellis Horwood Ltd.

- [18] Speetjens, J. (1976). Intermediate language for minicomputer cross-compilation.
- [19] Strong, J., Wegstein, J., Tritter, A., Olsztyn, J., Mock, O., and Steel, T. (1958).
The problem of programming communication with changing machines: a proposed solution.