עבודת גמר

לקבלת תואר טכנאי תוכנה

הנושא: **קומפיילר**

המגיש : איתן רפאל צ'רטוף

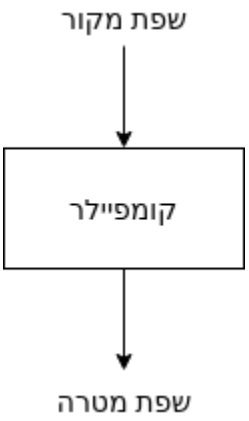
ת.ז. המגיש: 215310715

שמות המנחים : מיכאל

אפריל 2024 תשפ"ד

תקציר

שפת תכנות היא קבוצה של סימונים המשמשת לכתיבת תוכניות מחשב. העולם היום תלוי על שפות תכנות, מכיוון שכל התוכניות הרצות בעולם נכתבו באחת מן כל שפות התכנות. אך, בשביל להריץ תוכנית הכתובה משפת תכנות, צריך לתרגם אותה לצורה שהמחשב יכול להריץ.
המערכות שיכולות לעשות תרגום שכזה נקראות *קומפיילרים* (או בעברית תקינה, *מהדרים*. במהלך כל התיק אני אשתמש במילה קומפיילר). במילים פשוטות, קומפיילר הינו תוכנית המקבלת תוכנית הכתובה בשפת תכנות - *שפת המקור* - ומתרגם את התוכנית הזאת לתוכנית מקבילה לתוכנית המקורית, רק בשפה אחרת - *שפת המטרה*.

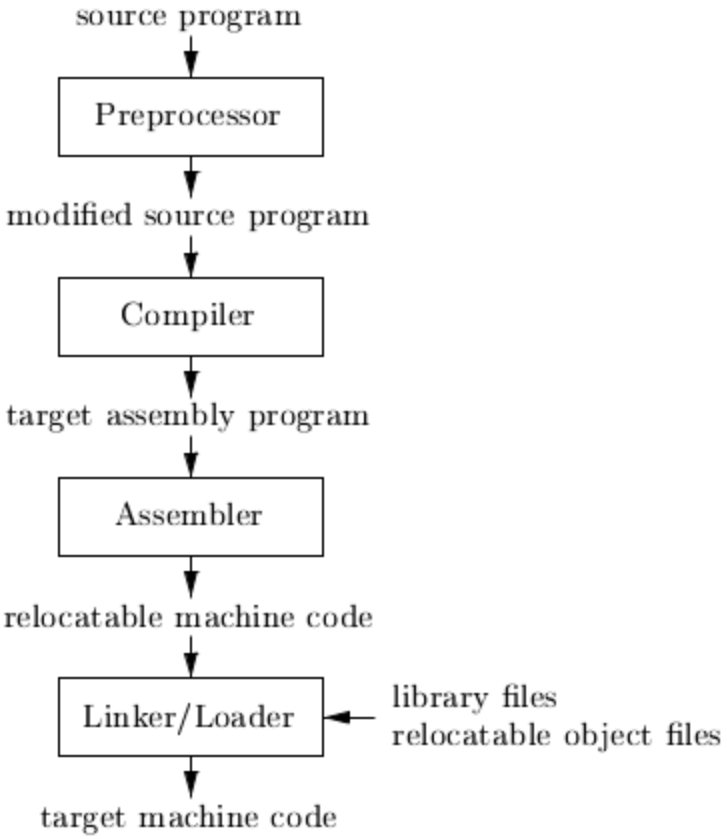


1.1. קומפיילר

עיבוד שפה

כאשר מעבדים שפת תכנות לשפת מכונה, הקומפיילר הוא רק חלק מהתהליך. עוד חלקים מהתהליך הם:

- מעבד מקדים - תכנית הקולטת נתונים מקדימים בשביל שהפלט שלה ישמש בתכנית אחרת. סוג תכנית זו תקרא תמיד לפני תכנית אחרת שתשתמש בפלו תכנית זאת, לכן השם עיבוד מקדים.
- מעבד שפת שף - מתרגם שפת סף לשפת מכונה
- מקשר - תכנית המחברת תוכניות מחשב שעברו הידור לשפת מכונה לתוכנית אחת.



1.2. מערכת עיבוד שפה

מטרת הקומפיילר בתהליך היא לקחת את הפלט של המעבד המקדים (שלא אמור להיות שונה בצורה גדולה מקוד המקור), ולתרגם אותו לשפת סף.

מושגים

קומפיילר/מהדר

תוכנית המקבלת תוכנית הכתובה בשפת תכנות - *שפת המקור* - ומתרגם את התוכנית הזאת לתוכנית מקבילה לתוכנית המקורית, רק בשפה אחרת - *שפת המטרה*.

מתורגמן/Interpreter

תכנית המבצעת ישירות הוראות שנכתבו בשפות תכנות מבלי לדרוש שהן תורגמו לשפת מכונה. בדרך כלל האינטרפרטר כולל קבוצה של הוראות שאפשר לבצע ורשימה של הוראות אלו לפי הסדר שהתכניתן רצה שההוראות יפעלו.
האינטרפרטר מתרגם ומבצע את התכנית הרצויה שורה אחרי שורה, לכן בדרך כלל האינטרפרטרים יהיו איטיים מקומפיילרים, המתרגמים את כל התכנית.

מושגי תכנות

משתנים

מקום אחסון בעל שם וסוג ערך שמור. אפשר להתייחס למשתנים בעזרת שמם או כתובת הזיכרון שלהם. בזמן ריצת תכנית מחשב אפשר להכין משתנים, להגדיר להם ערך, לשנות ערך זה, למחוק את המשתנים ועוד.
דוגמאות להגדרת משתנים שישמשו כמידע על בן אדם:

```
// גיל המוגדר כמספר שלם
int age;
// שם המוגדר כמחרוזת של אותיות
string name;
// מספר אהוב המוגדר כשבר
float fav_number;
```

תנאים

הוראה הבודקת תנאי מסוים. תנאים הם דרך לבדוק תנאי מסוים בזמן ריצת התכנית, ואפשר להשתמש במשתנים בתנאים. תנאים בדר"כ כתובים בהוראות if - else, הכוונה היא אם קורה משהו, תעשה משהו, ואם לא תעשה משהו אחר:

```
if (condition):
  statement
else:
  statement
```

לולאות

לולאות משומשות בשביל להריץ חלק של הקוד שוב ושוב עד שתנאי מסוים מתקיים. יש שני סוגים עיקריים של לולאות, ה - for loop וה - while loop. בדר"כ משתמשים בfor שאנחנו יודעים את מספר האיטרציות, ובwhile שאנחנו לא. דוגמא לשני תוכניות המדפיסות 10 כוכביות, אחת לאחר השנייה:

```
int i;
for(i = 0; i < 10; ++i) {
  printf("*");
}

int i = 0;
while(i < 10) {
  printf("*");
  ++i;
}
```

ארכיטקטורת הפתרון

1. ניתוח מילוני (lexical analysis)

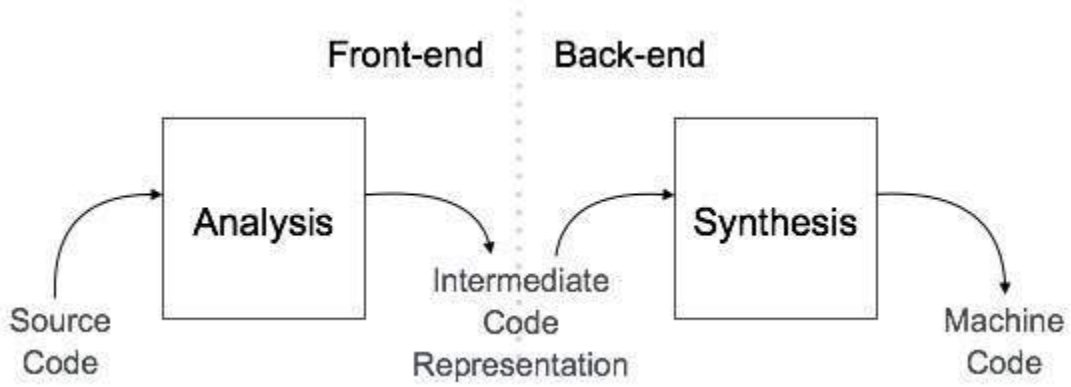
בחלק זה התכנית קוראת את הקוד והופכת אותו אל אסימונים או "lexical tokens", שהם כמו היחידות הבסיסיות של השפה. אפשר לקרוא לתכנית זו ה"lexer". החלק הזה מבטיח שכל המילים הנמצאות בשפה מותאמות למילון השפה.

2. ניתוח תחביר (syntax analysis/parsing)

חלק זה קולט את האסימונים שקיבלנו מהתכנית הקודמת ומחיל אליהם כללים מוגדרים כדי לקבוע עם הקוד נכון מבחינה תחבירית. קוראים לתכנית זו ה"parser". המנתח התחבירי מבטיח שאין לקוד המקור בעיה אם התחביר.

3. ניתוח סמנטי (semantic analysis)

בשלב זה התכנית בוחנת האם לקוד יש שגיאות סמנטיות כמו טיפוסים לא תואמים או משתנה שלא הוגדר. השלב הזה מבטיח שאין שגיאות סמנטיות בקוד המקור.



4. יצירת קוד ביניים (intermediate code generation)

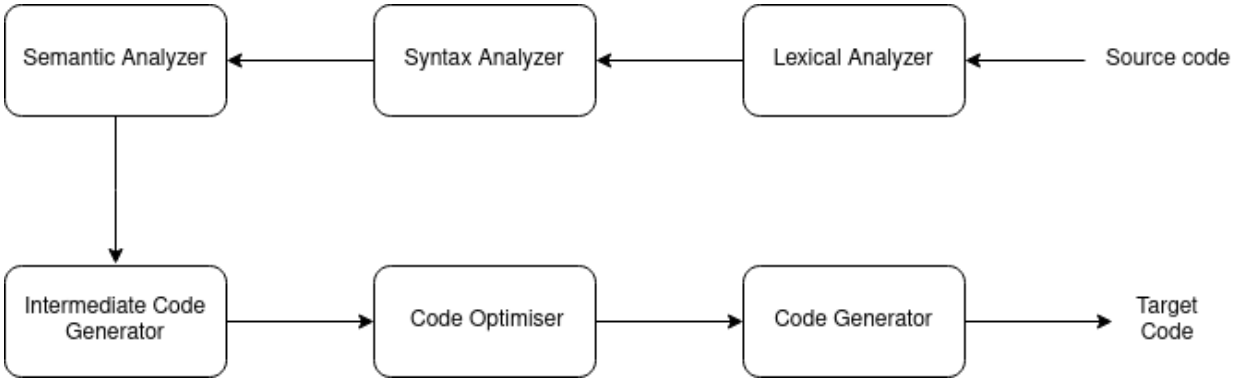
לאחר ניתוח הקוד, התכנית מייצרת פרזנטציה שונה של הקוד ופולטת אותה בשביל שהקומפיילר אשתמש בא לאופטימיזציה ותרגום.

5. אופטימיזציה (code optimization)

שלב זה משתמש בפלט של התכנית הקודמת ומשפר אותה. תכנית זו עושה דברים כמו מוחקת שורות קוד מיותרות ומסדרת את ההוראות בשביל למהר התכנית הסופית.

6. יצירת קוד (code generation)

בשלב זה התכנית קולטת את הקוד המשופר ומתרגמת אותו לשפת המחשב הרצויה.



כל רכיב זה הוא מבנה משל עצמו עם קלט ופלט שונה, והם משתמשים כמו שכבות בשביל להגיע לתוצאה הסופית, קוד מקומפל.

Symbol table

טבלת סמלים היא מבנה נתונים המשמש את המהדר לאחסון מידע על הסמלים השונים, כגון מזהים, קבועים, נהלים ופונקציות, בקוד המקור של תוכנית. היא שומרת מידע חיוני על כל סמל, כולל שמו, סוגו, תחומו (scope) ותכונות אחרות שלו. טבלת הסמלים משומשת במהלך כל שלבי הקומפילציה, ונבנת בשלבי הניתוח. מטרות טבלת הסמלים:

- זיהוי סוג משתנים, כתובת הזיכרון ושמות של משתנים.
- ניהול משתנים בהתייחס לתחום.
- משומש בשביל להתמודד עם שגיאות.
- הטבלה מסדרת את הסמלים והתכונות שלהם מה שעוזר לניהול התכנית.
- משתמשים בטבלה בשביל ליצור את הקוד הסופי.

התמודדות עם שגיאות

בנוסף לכל הרכיבים האלו, הקומפיילר צריך לדעת איך להתמודד עם שגיאות. לכל מבנה שונה יהיה שגיאות שונות, לדוגמא, במנתח המילולי, יכול להיות שגיאה לקסיקלית, שבא אותו מנתח מזה רצף תווים לא מזוהה ואינה יודעת איך להתמודד אם אותו רצף. נחלק את השגיאות שיכולות להיות לנו לארבעה קטגוריות:

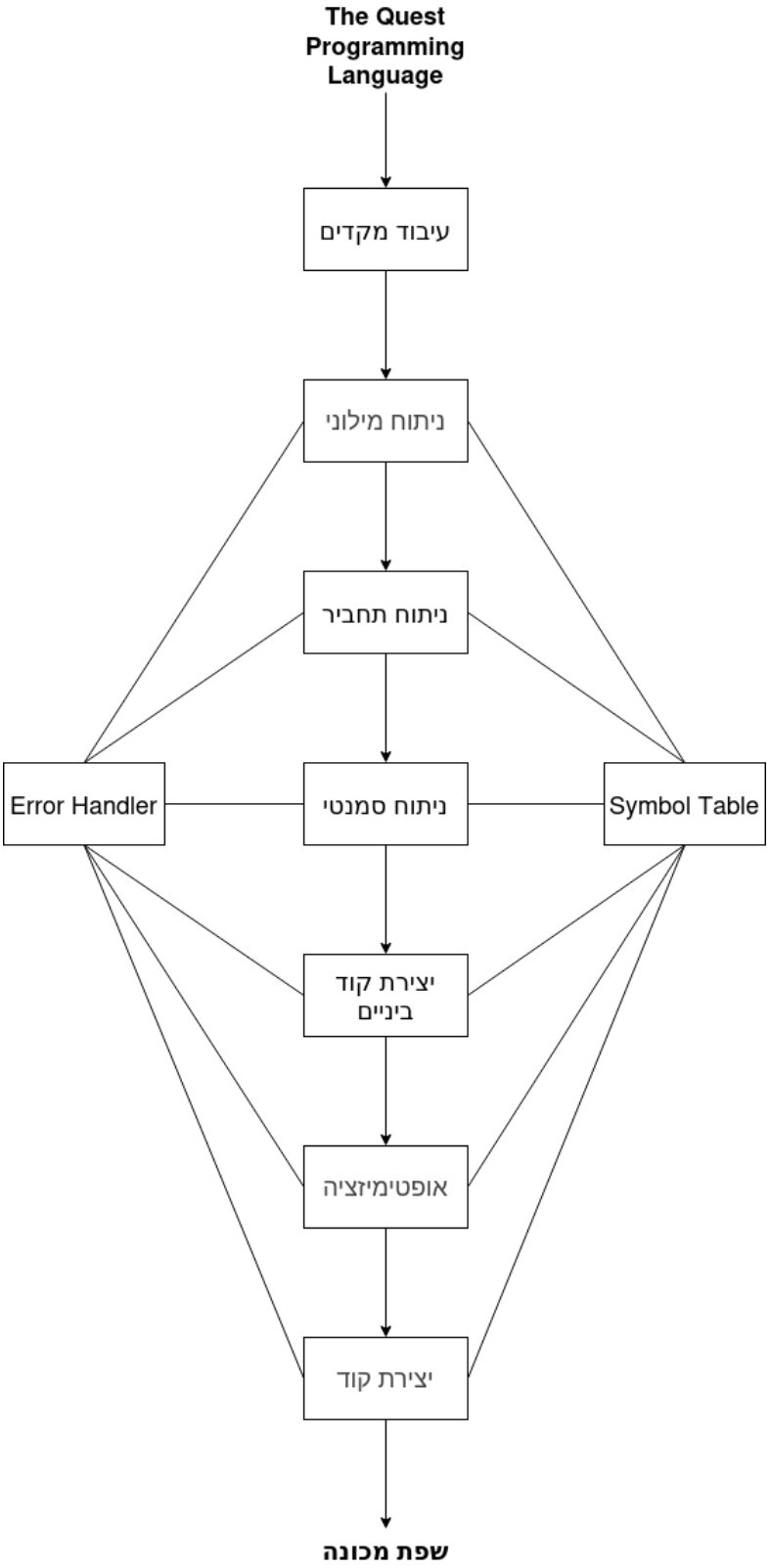
1. **שגיאות לקסיקליות** - המנתח המילוני קורא רצף תווים לא מזוהה ולא יודע איך להתמודד עם אותו הרצף, לדוגמה, שם של סוג משתנה רשום בצורה שגויה.
2. **שגיאות תחביריות** - המנתח התחבירי מזהה בתחביר החומרה, כלומר אסימון במקום שהוא לא צריך להיות, לדוגמא, סוג משתנה במקום לא רצוי או סוגריים פותחות ללא סוגר מתאים.
3. **שגיאות סמנטיות** - המנתח הסמנטי מזהה שגיאה סמנטית, לדוגמא חיבור בין שני סוגי משתנה שונים בלי דרך למצוא תוצאה תואמת.
4. **שגיאות לוגיות** - שגיאות בלוגיקה של הקוד, לדוגמא לולאה אין-סופית.

בשביל לזהות את אותם שגיאות, נשתמש בעוד רכיב הנקרא **מטפל השגיאות (Error Handler)**.
למטפל השגיאות יהיה שלושה ייעודים:

- זיהוי שגיאות

- דיווח שגיאות (במידה ורצוי)
- טיפול בשגיאות (במידה ואפשר)

תרשים סופי


The Quest Programming Language
עיבוד מקדים
ניתוח מילוני
ניתוח תחביר
Error Handler
ניתוח סמנטי
Symbol Table
יצירת קוד ביניים
אופטימיזציה
יצירת קוד
שפת מכונה

מבנה נתונים

במהלך פרק זה יתוארו כל מבני הנתונים, אציין אם הם כלליים בשביל כל שלבי הקומפילציה או שהם מיוחדים לשלב מסויים.

חשוב לזכור למה אנחנו בוחרים מבני נתונים אחד לגבי השני, אפילו שבסופו של דבר הם יעזרו לנו להגיע לאותה מטרה. מבני נתונים הופכים את האלגוריתמים שלנו ליותר יעילים, מפחיתים את זמן הריצה, חוסכים בזיכרון ויכולים להפחית את המורכבות והסיבוכיות של אלגוריתם ספציפי, ומבני נתונים מסוימים עושים את העבודה הזאת יותר טוב ממבני נתונים אחרים. כמובן שבחירת מבני הנתונים הכי טוב היא תלות האלגוריתם.

מבני נתוני המנתח המילוני

אוטומט דטרמיניסטי סופי (Deterministic finite automaton)

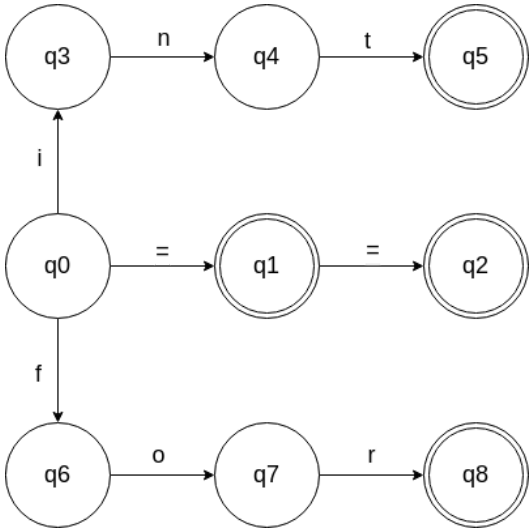
המנתח המילולי קורא את קוד המקור תו אחר תו בשביל להתאים את המילים לאסימונים המתאימים שלהם, לכן חשוב שההתאמה הזאת תהיה כמה שיותר אופטימלית. כלומר, שסיבוכיות הזמן של אלגוריתם ההתאמה בין מילה לאסימון תהיה בעל החסם האסימפטוטי העליון הקטן ביותר. לכן, בשביל שההתאמה תהיה כמה שיותר אופטימלית, בחרתי להשתמש ב*אוטומט דטרמיניסטי סופי*, ונשתמש בו להכין אלגוריתם שמתאים מילה לאסימון במילון השפה בסיבוכיות O(1) .

אוטומט דטרמיניסטי סופי (באנגלית Deterministic finite automaton או DFA בקיצור) הוא מודל מתמטי המגדיר שפה פורמלית. המודל מורכב מקבוצה סופית של מצבים בעל כללי מעבר ביניהם, כלומר חוקים המגדירים מה לעשות במצב כאשר נקלט אות מהקלט. הקלט יהיה אוסף של אותיות (מילה) מתוך הא"ב של השפה, שהיא בנויה מקבוצה של סימנים, לדוגמה, בשפת Quest, הא"ב יהיה כל תווי הascii. הסיבה שהאוטומט דטרמיניסטי היא שכל מצב ידוע מראש ומוגדר.

נממש מבנה זה בשפת C בעזרת מערך דו-ממדי, כאשר:

- כל 128 תווי ה-ascii יוצגו ע"י עמודות המערך הדו-ממדי, כאשר כל אינדקס מותאם לערך האות (לפי טבלת הascii).
- כל המצבים ייוצגו ע"י שורות המערך הדו-ממדי, המצב האחרון שנהיה בוא יהיה תואם לאסימון.
  - אנחנו נגדיר מצב מתחיל שבו תמיד נתחיל את האנליזה, שהוא יהיה המצב הראשון שנקרא לו q0. המצב השני יהיה q1, השלישי q2 וכך הלאה...
- כל ערך בעמודה התואמת לסימן ובשורה התואמת למצב יהיה שווה למצב הבא שהאוטומט יקפוץ אליו.
- לכל אות במצב שאנחנו יודעים שלא מתאימה, נגדיר ערך 1- במקומה, בשביל להגדיר שהמילה אינה במילון השפה.

לדוגמא, נגדיר DFA שלוקח את המילים: int, for, ==, =



נמיר את זה למערך דו ממדי (נתעלם מזה שאינדקס העמודה מתאים לערך האות):

	=	i	n	t	f	o	r
0	1	3	-1	-1	6	-1	-1
1	-1	2	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	4	-1	-1	-1	-1
4	-1	-1	-1	5	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	7	-1
7	-1	-1	-1	-1	-1	-1	8
8	-1	-1	-1	-1	-1	-1	-1