תורת הקומפילציה

תרגיל 3

ההגשה בזוגות בלבד

שאלות במייל אל האחראי על התרגיל בלבד bahjat.kawar@campus.technion.ac.il

לתרגיל יפתח דף FAQ באתר הקורס. יש להתעדכן בו. הנחיות והבהרות שיופיעו בדף ה-FAQ עד יומיים לפני הגשת התרגיל מחייבות. שאלות במייל המופיעות בדף ה-FAQ לא יענו.

התרגיל ייבדק בבדיקה אוטומטית. **הקפידו למלא אחר ההוראות במדויק.**

כללי

בתרגיל זה עליכן לממש ניתוח תחבירי לשפת FanC, הכוללת פעולות אריתמטיות, enums, פונקציות, והמרות מובנות מ-byte (בית אחד) ל-int (4 בתים).

מנתח לקסיקלי

יש לכתוב מנתח לקסיקלי המתאים להגדרות הבאות:

אסימון	תבנית
VOID	void
INT	int
BYTE	byte
B	b
BOOL	bool
ENUM	enum
AND	and
OR	or
NOT	not
TRUE	true
FALSE	false
RETURN	return
IF	if
ELSE	else
WHILE	while
BREAK	break
CONTINUE	continue
SC	;
COMMA	,
LPAREN	(
RPAREN	)
LBRACE	{
RBRACE	}
ASSIGN	=
RELOP	== | != | < | > | <= | >=
BINOP	+ | - | * | /
ID	[a-zA-Z][a-zA-Z0-9]*
NUM	0 | [1-9][0-9]*
STRING	"([^\n\r\"\\]|\\[rnt"\\])+"

ניתן לשנות את שמות האסימונים או להוסיף אסימונים נוספים ,כל עוד המנתח הלקסיקלי מזהה את כל התבניות לעיל.
יש להתעלם מרווחים, ירידות שורה משני הסוגים (LF ,CR) וטאבים כך שלא תתקבל עליהם שגיאה לקסיקלית.
יש להתעלם מהערות שורה (הערות ++C) המיוצגות ע"י התבנית //[^\r\n]*[\r|\n|\r\n]?

תחביר

יש לכתוב מנתח תחבירי שיתאים לדקדוק הבא:

1. $Program \rightarrow Enums\ Funcs$
2. $Funcs \rightarrow \epsilon$
3. $Funcs \rightarrow FuncDecl\ Funcs$
4. $FuncDecl \rightarrow RetType\ ID\ LPAREN\ Formals\ RPAREN\ LBRACE\ Statements\ RBRACE$
5. $Enums \rightarrow \epsilon$
6. $Enums \rightarrow EnumDecl\ Enums$
7. $EnumDecl \rightarrow ENUM\ ID\ LBRACE\ EnumeratorList\ RBRACE\ SC$
8. $RetType \rightarrow Type$
9. $RetType \rightarrow VOID$
10. $Formals \rightarrow \epsilon$
11. $Formals \rightarrow FormalsList$
12. $FormalsList \rightarrow FormalDecl$
13. $FormalsList \rightarrow FormalDecl\ COMMA\ FormalsList$
14. $FormalDecl \rightarrow Type\ ID$
15. $FormalDecl \rightarrow EnumType\ ID$
16. $EnumeratorList \rightarrow Enumerator$
17. $EnumeratorList \rightarrow EnumeratorList\ COMMA\ Enumerator$
18. $Enumerator \rightarrow ID$
19. $Statements \rightarrow Statement$
20. $Statements \rightarrow Statements\ Statement$
21. $Statement \rightarrow LBRACE\ Statements\ RBRACE$
22. $Statement \rightarrow Type\ ID\ SC$
23. $Statement \rightarrow EnumType\ ID\ SC$
24. $Statement \rightarrow EnumDecl$
25. $Statement \rightarrow Type\ ID\ ASSIGN\ Exp\ SC$
26. $Statement \rightarrow EnumType\ ID\ ASSIGN\ Exp\ SC$
27. $Statement \rightarrow ID\ ASSIGN\ Exp\ SC$
28. $Statement \rightarrow Call\ SC$
29. $Statement \rightarrow RETURN\ SC$
30. $Statement \rightarrow RETURN\ Exp\ SC$
31. $Statement \rightarrow IF\ LPAREN\ Exp\ RPAREN\ Statement$
32. $Statement \rightarrow IF\ LPAREN\ Exp\ RPAREN\ Statement\ ELSE\ Statement$
33. $Statement \rightarrow WHILE\ LPAREN\ Exp\ RPAREN\ Statement$
34. $Statement \rightarrow BREAK\ SC$
35. $Statement \rightarrow CONTINUE\ SC$
36. $Call \rightarrow ID\ LPAREN\ ExpList\ RPAREN$
37. $Call \rightarrow ID\ LPAREN\ RPAREN$
38. $ExpList \rightarrow Exp$

39. $ExpList \rightarrow Exp\ COMMA\ ExpList$
40. $Type \rightarrow INT$
41. $Type \rightarrow BYTE$
42. $Type \rightarrow BOOL$
43. $EnumType \rightarrow ENUM\ ID$
44. $Exp \rightarrow LPAREN\ Exp\ RPAREN$
45. $Exp \rightarrow Exp\ BINOP\ Exp$
46. $Exp \rightarrow ID$
47. $Exp \rightarrow Call$
48. $Exp \rightarrow NUM$
49. $Exp \rightarrow NUM\ B$
50. $Exp \rightarrow STRING$
51. $Exp \rightarrow TRUE$
52. $Exp \rightarrow FALSE$
53. $Exp \rightarrow NOT\ Exp$
54. $Exp \rightarrow Exp\ AND\ Exp$
55. $Exp \rightarrow Exp\ OR\ Exp$
56. $Exp \rightarrow Exp\ RELOP\ Exp$
57. $Exp \rightarrow LPAREN\ Type\ RPAREN\ Exp$

הערות:

1. הדקדוק כפי שמוצג כאן אינו חד משמעי ב-Bison. יש להפכו לחד משמעי תוך שימור השפה. בעיה לדוגמה שיש לפתור: http://en.wikipedia.org/wiki/Dangling_else.
יש לפתור את בעיית ה-Dangling else (למשמעות כמו בשפת C) ללא שינוי הדקדוק אלא באמצעות מתן עדיפות לכללים או אסוציאטיביות מתאימה לאסימונים.

2. יש להקפיד על מתן עדיפויות ואסוציאטיביות מתאימים לאופרטורים השונים. יש להשתמש בטבלת העדיפויות כאן: http://introcs.cs.princeton.edu/java/11precedence

3. אין צורך לבצע שינויים בדקדוק, פרט לשם הבדלה בין האופרטורים השונים.

בדיקות סמנטיות

טבלאות סמלים

בשפת FanC קיים קינון סטטי של scopes: כל משתנה, טיפוס enum או ערך של טיפוס enum. מוגדר ב-scope שבו הוכרז, ובכל הצאצאים של אותו scope. אסור להכריז על משתנה, טיפוס או ערך שכבר מוגדר באותו ה-scope – כלומר **אין shadowing של אף identifier שכבר מוגדר** (כולל identifier של פונקציה), ואסור להשתמש במשתנה, פונקציה, טיפוס או ערך שלא הוגדרו. שימוש במשתנה הוא כל מופע פרט להכרזה שלו. משתנה מוגדר החל מה-statement שאחרי הגדרתו.

קטעי הקוד הבאים תקינים תחבירית:

```
int a;
int a;
```

וגם:

```
int a;
c = 6;
```

אך לא נרצה לאפשר אותם בשפת FanC. לכן יש לנהל טבלאות סמלים.

בטבלת הסמלים נשמור עבור כל משתנה, פרמטר ופונקציה את שמו, מיקומו היחסי ברשומת ההפעלה, והטיפוס שלו.

יש להשתמש בטבלאות הסמלים כדי לבצע את הבדיקות הבאות:

1. בכל הכרזה על משתנה יש לוודא שמשתנה באותו שם לא מוגדר ב-scope הנוכחי או באחד ה-scopes המכילים אותו.
2. בכל שימוש במשתנה יש לוודא כי הוא מוגדר.
3. בכל שימוש בפונקציה, יש לוודא כי היא הוגדרה לפני השימוש. כלומר: מותר לקרוא לכל פונקציה שהוגדרה לפני הפונקציה הנוכחית, ומותר לקרוא לפונקציה הנוכחית (רקורסיה).

בנוסף יש להשתמש בטבלת הסמלים כדי לבצע בדיקות טיפוסים לפי כללי הטיפוסים של FanC שמוגדרים בהמשך.

שימו לב כי בשביל לתמוך במשתנים מסוג enum ובפונקציות ייתכן שתצטרכו לשמור מידע נוסף פרט למידע לעיל.

הקינון הסטטי של ה-scopes רלוונטי גם עבור טיפוסי ה-enum אותם ניתן להגדיר בתוכנית. ניתן להגדיר משתנה מטיפוס enum <id> רק במידה שטיפוס זה הוגדר ב-scope הנוכחי או באב קדמון שלו (כולל הscope הגלובלי). **ניתן להעביר טיפוס מסוג enum <id> כארגומנט לפונקציה רק במידה שהוא הוגדר ב-scope הגלובלי.**

כללי Scoping:

1. פונקציה ובלוק מייצרים scope חדש. פרמטרים של פונקציה שייכים ל-scope של הפונקציה.
2. if/else/while מייצרים scope חדש.

לכן, במקרה בו נפתח בלוק כחלק מפקודת if/else/while יפתחו שני scopes. אחד ריק עבור ה-if/while/else ואחד עבור הבלוק.

בנוסף קיימות שתי פונקציות ספריה: print ו-printi, כאשר print מקבלת מחרוזת (string) ו-printi מקבלת int. שתיהן מחזירות void. יש להכניס את שתי הפונקציות הנ"ל לטבלת הסמלים בפתיחת הscope הגלובלי בסדר הבא: קודם את print ולאחר מכן את printi.

שימו לב כי כדי לשמור את print בטבלת הסמלים אנחנו מגדירים את string כטיפוס פנימי, למרות שהוא לא נגזר ע"י Type.

כללי טיפוסים

יש לקבוע את הטיפוסים של ביטויים לפי הכללים הבאים:

1. ביטוי שנגזר מ-NUM טיפוסו int, ומ-NUM B טיפוסו byte.
2. טיפוס הקבועים true/false הוא bool.
3. טיפוס קבוע מחרוזת הוא string.
4. הטיפוס של משתנה נקבע לפי הגדרתו.
5. כל הגדרת enum עם <id> מגדירה טיפוס מסוג enum <id>.
6. הערכים שמשתנה מטיפוס enum <id> יכול לקבל הינם רק הערכים שהוגדרו בעת הגדרת הטיפוס.
7. הטיפוס של ביטוי Call נקבע לפי טיפוס ההחזרה של הפונקציה הנקראת.
8. ניתן לבצע השמה של ביטוי מטיפוס מסוים למשתנה מאותו הטיפוס.
9. ניתן לבצע השמה של byte ל-int.
10. ניתן לבצע המרה מפורשת של משתנה <var> מטיפוס enum <id> ל-int **בלבד** ע"י הפקודה: (int) <var>.
כאשר מבצעים המרה זו הערך המספרי של <value> הוא המספר הסידורי של הערך לפי סדר הגדרת הערכים לטיפוס (כאשר הספירה מתחילה מ-0).
לדוגמה:

```
enum day {SUNDAY, MONDAY, TUESDAY, WEDNESDAY, THURSDAY, FRIDAY,
SATARDAY};
enum day d = TUESDAY;
```

```
printi((int) d);  // prints 2
```

11. פעולות relop מקבלות שני אופרנדים מטיפוסים מספריים. טיפוס ההחזרה של הביטוי הוא bool.
12. פעולות לוגיות (and, or, not) מקבלות אופרנדים מטיפוס bool. טיפוס ההחזרה של הביטוי הוא bool.
13. פעולות binop מקבלות שני אופרנדים מספריים. טיפוס החזרה של binop הוא הטיפוס עם טווח הייצוג הגדול יותר מבין שני הטיפוסים של האופרנדים.
14. ביטוי מסוג string ניתן לשימוש רק בקריאה לפונקציית הספרייה print.
15. פונקציית הספריה print מקבלת ארגומנט אחד מסוג string ומחזירה void.
16. פונקציית הספרייה printi מקבלת ארגומנט אחד מסוג int או byte ומחזירה void.
17. ניתן לקרוא לפונקציה בהעברת מספר נכון של פרמטרים תואמים לטיפוסים בהגדרת הפונקציה (לפי הסדר). מותר להעביר ביטוי e_i לפרמטר p_i של הפונקציה אם השמה של e_i למשתנה המוגדר מהטיפוס של p_i מותרת.
18. באותו אופן, בפונקציה המחזירה ערך, טיפוס ה-Exp בכל $RETURN\ Exp$ חייב להיות מותר להשמה לטיפוס ההחזרה בהגדרת הפונקציה.
19. פקודות if ו-while מקבלות Exp מטיפוס בוליאני.

שימו לב! בכל מקרה שלא מוגדר בכללים אלה יש להחזיר שגיאה. ראו סעיף "טיפול בשגיאות" בהמשך.

בדיקות סמנטיות נוספות

בנוסף, יש לבצע את הבדיקות הבאות, שאינן בדיקות טיפוסים:

1. עבור כלל הדקדוק $Statement \rightarrow BREAK\ SC$ ועבור הכלל $Statement \rightarrow CONTINUE\ SC$ יש לבצע בדיקה כי הם מתגלים רק בתוך לולאת while, אחרת יש לעצור עם שגיאת UnexpectedBreak או UnexpectedContinue בהתאמה ולצאת מהתכנית.
2. עבור כללי הדקדוק $Statement \rightarrow RETURN\ SC$ ו-$Statement \rightarrow RETURN\ Exp\ SC$ יש לבצע בדיקה כי הם תואמים לטיפוס הפונקציה: $RETURN\ Exp\ SC$ מותר לשימוש רק בפונקציות שלא מחזירות void (בדיקת הטיפוס עבורו מפורטת תחת "כללי טיפוסים"), ו-$RETURN\ SC$ רק בפונקציה המחזירה void. אחרת יש לעצור עם שגיאת Mismatch ולצאת מהתכנית.
   שימו לב שאין חובה שפונקציה תכיל פקודת return ואין צורך לבדוק שלפונקציה המחזירה ערך קיימת פקודת return.
3. ליטרל שטיפוסו byte לא יציין מספר הגדול מ-255.
4. קיימת בדיוק פונקציית main אחת, ללא פרמטרים, ועם טיפוס החזרה void.

מיקום המשתנים בזיכרון

בתרגיל אנו מניחים שכל משתנה הוא בגודל 1, ללא תלות בטיפוס. אזי עבור הקוד הבא:

```
int x;
{
     bool y;

     byte z;
}
bool w;
```

המיקומים (offset) לכל משתנה יהיו:

x	0
y	1
z	2
w	1

בנוסף, נמקם ארגומנטים של פונקציה בסדר הפוך ברשומת ההפעלה לפני מיקום 0. לכן עבור הפונקציה הבאה:

```
bool isPassing(int grade, int factor)
{
      return (grade+factor) > 55;
}
```

המיקומים יהיו:

grade	-1
factor	-2

קלט ופלט המנתח

קובץ ההרצה של המנתח יקבל את הקלט מ-stdin.

יש להיעזר בקובץ output.hpp המצורף לתרגיל על מנת לייצר פלט הניתן לבדיקה אוטומטית.

בסוף כל scope, כולל ה- scope הגלובאלי, המנתח ידפיס את המשתנים שהוגדרו ב- scope זה ל-stdout **בסדר הבא**:

1. קריאה לפונקציה endScope
2. עבור כל identifier (שאינו הגדרת enum או ערך של enum) שהוגדר ב- scope על פי סדר ההכרזה בקוד (במידה ומדובר ב-scope של פונקציה, יש להתחיל מהפרמטרים, לפי סדר הגדרתם) יש לקרוא לפונקציה printID(id,offset,type) עם שם המשתנה, המיקום בזיכרון, והטיפוס.
   a. עבור משתנה או פרמטר שאינם enum, מחרוזת הטיפוס צריכה להיות זהה לשם האסימון שהוגדר לטיפוס בחלק הלקסיקלי בתיאור התרגיל (עבור מחרוזת, הטיפוס הוא STRING).
   b. עבור משתנה או פרמטר מטיפוס enum <id>, מחרוזת הטיפוס צריכה להיות enum <id>.
   c. עבור פונקציה, כדי לקבל את מחרוזת הtype יש לקרוא לפונקציה makeFunctionType עם טיפוסי הפרמטרים וטיפוס ההחזרה כפי שהוגדרו בסעיף הקודם. בנוסף, המיקום בזיכרון של פונקציה הוא תמיד 0.
3. עבור כל טיפוס enum שהוגדר ב-scope על פי סדר ההכרזה בקוד יש לקרוא לפונקציה printEnumType עם שם ה-enum והערכים שלו לפי הסדר בו הם הוגדרו.
4. שימו לב לבצע זאת בסוף כל scope לפי ההגדרה בפרק טבלת הסמלים של תיאור התרגיל.

ניתן קובץ פלט לדוגמא. יש לבדוק שהפורמט שהודפס זהה אליו. הבדלי פורמט יגרמו לכישלון הבדיקות האוטומטיות.

טיפול בשגיאות

בקובץ הקלט יכולות להיות שגיאות לקסיקליות, תחביריות וסמנטיות. **על המנתח לסיים את ריצתו מיד עם זיהוי שגיאה** (כלומר בנקודה העמוקה ביותר בעץ הגזירה שבה ניתן לזהותה). ניתן להניח כי הקלט מכיל שגיאה אחת לכל היותר.

על מנת לדווח על שגיאות יש להשתמש בפונקציות הנתונות בקובץ output.hpp:

שגיאה לקסיקלית	errorLex(lineno)
שגיאה תחבירית	errorSyn(lineno)
שימוש במשתנה שלא הוגדר או ב-identifier שאינו משתנה כמשתנה	errorUndef(lineno, id)
שימוש בפונקציה שלא הוגדרה או ב-identifier שאינו פונקציה כפונקציה	errorUndefFunc(lineno, id)
ניסיון להגדיר identifier שכבר הוגדר	errorDef(lineno, id)
ניסיון להשתמש בפונקציה עם ארגומנטים לא תואמים. types יהיה רשימת הטיפוסים המצופים.	errorPrototypeMismatch(lineno, id, types)
השמת ערך לא חוקי למשתנה/פרמטר id מטיפוס enum.	errorUndefEnumValue(lineno, id)

אי התאמה של טיפוסים (פרט להעברת פרמטרים לא תואמים לפונקציה ולהשמת ערך לא חוקי למשתנה/פרמטר id מטיפוס enum).	errorMismatch(lineno)
פקודת break שאינה חלק מלולאה	errorUnexpectedBreak(lineno)
פקודת continue שאינה חלק מלולאה	errorUnexpectedContinue (lineno)
לא מוגדרת פונקציית void main()	errorMainMissing()
ליטרל מסוג byte מכיל מספר גדול מדי, כאשר value הוא הערך הקיים בקוד.	errorByteTooLarge(lineno, value)
הצהרה על משתנה/פרמטר id מטיפוס enum שלא הוגדר.	errorUndefEnum(lineno, id)

בכל השגיאות הנ"ל id הוא שם המשתנה או הפונקציה, ו-lineno הוא מס' השורה בה מופיעה השגיאה.

- במקרה של שגיאה הפלט של המנתח יהיה תוכן כל ה-scopes שנעשה להם reduce והשגיאה שהתגלתה (כפי שניתן לראות בדוגמה t2).
- יש לתפוס את השגיאה ולעצור את המנתח מוקדם ככל הניתן.
לדוגמה, במקרה שבתנאי if מופיע Exp שאינו מטיפוס בוליאני, יש לזרוק את השגיאה ולעצור לפני ההדפסה שמתבצעת בסוף הscope.
- בדיקה כי קיימת פונקציית void main() תתבצע לפני reduce של ה-scope הגלובלי . ולכן על המנתח לזהות זאת לפני הדפסת תוכן ה-scope הגלובלי.

הדרכה והערות

סדר מומלץ לביצוע התרגיל (מומלץ להריץ בדיקות לאחר כל סעיף):

1. כתבו מנתח לקסיקלי ותחבירי ללא כללים סמנטיים.
2. בדקו שהמבנה התחבירי של השפה נאכף ואין אף קונפליקט.
3. הגדירו את YYSTYPE וממשו טבלאות סמלים. השתדלו ליצור מחלקות לכל נונטרמינל ולא ליצור struct אחד שמכיל את כל התכונות הסמנטיות.
מלבד הצורה שראיתם בתרגול לעשות זאת, ניתן לעשות זאת גם ע"י הגדרת union המכיל את כל ה-structs או הטיפוסים האפשריים. והגדרת כל טרמינל ונונטרמינל כ-struct או טיפוס המתאים לו.
להסבר ולדוגמה פשוטה עבור דקדוק המכיל טרמינלים NUM ו-OP ונונטרמינל exp נוסיף בחלק ה-declarations:

```
%union {
int val;
char op;
};
%token <val> NUM
%token <op> OP
%type  <val> exp
```

4. מומלץ מאוד לממש מחלקות לטיפול בדרישות שונות ולהפנות אליהן מהקוד בקובץ הדקדוק. שימוש בקוד חיצוני יחסוך לכם להריץ את bison בכל שינוי של הקוד. שימו לב כי ניתן להגיש קבצי קוד נוספים.
5. בצעו בדיקות סמנטיות.

שימו לב כי התרגיל לא ייבדק עם הכלי valgrind. על אף זאת, על התרגיל לא לקרוס.

הוראות הגשה

שימו לב כי קובץ ה-Makefile מאפשר שימוש ב-STL. אין לשנות את ה-Makefile.

יש להגיש קובץ אחד בשם ID1-ID2.zip, עם מספרי ת"ז של שתי המגישות. על הקובץ להכיל:

- קובץ flex בשם scanner.lex המכיל את כללי הניתוח הלקסיקלי
- קובץ בשם parser.ypp המכיל את המנתח
- את כל הקבצים הנדרשים לבניית המנתח, כולל *.hw3_output שסופקו כחלק מהתרגיל.

בנוסף, יש להקפיד שהקובץ **לא יכיל** את:

- קובץ ההרצה
- קבצי הפלט של flex ו-bison
- קובץ Makefile שסופק כחלק מהתרגיל

יש לוודא כי בביצוע unzip לא נוצרת תיקיה נפרדת. על המנתח להיבנות על השרת csComp ללא שגיאות באמצעות קובץ Makefile שסופק עם התרגיל. באתר הקורס מופיע קובץ zip המכיל קבצי בדיקה לדוגמה. יש לוודא כי פורמט הפלט זהה לפורמט הפלט של הדוגמאות הנתונות. כלומר, ביצוע הפקודות הבאות:

```
unzip id1-id2.zip
cp path-to/Makefile .
cp path-to/ hw3-tests.zip .

unzip hw3-tests.zip
make
./hw3 < t1.in 2>&1 > t1.res
diff t1.res path-to/t1.out
```

ייצור את קובץ ההרצה בתיקיה הנוכחית ללא שגיאות קומפילציה, יריץ אותו, ו-diff יחזיר 0.

הגשות שלא יעמדו בדרישות לעיל יקבלו ציון 0 ללא אפשרות לבדיקה חוזרת.

בדקו היטב שההגשה שלכן עומדת בדרישות הבסיסיות הללו לפני ההגשה עצמה.

שימו לב כי באתר מופיע script לבדיקה עצמית לפני ההגשה בשם selfcheck. תוכלו להשתמש בו על מנת לוודא כי ההגשה שלכם תקינה.

בתרגיל זה (כמו בתרגילים אחרים בקורס) **ייבדקו העתקות**. אנא כתבו את הקוד שלכם בעצמכם.

בהצלחה! ☺