אוניברסיטת חיפה - החוג למדעי המחשב
בלשנות חישובית, סמסטר א' תשס"ד

Introduction to PC-PATR program

The pc-patr program is free and available at http://www.sil.org/pcpatr/. It is available for MS-DOS, Microsoft Windows, Macintosh, and Unix.

אנו נשתמש בתוכנה שימוש מוגבל – עבור ניתוח והרצת דקדוקים חסרי הקשר.
התוכנה קולטת דקדוק חסר הקשר בצורה נורמלית (normal form) בשני קבצים – קובץ הלקסיקון וקובץ הדקדוק, ולאחר מכן מבצעת ניתוח למחרוזות שונות לבחירתנו כולל מתן כל עצי הגזירה האפשריים.

הערה: ניתן ע"י מניפולציות שונות להכניס כקלט גם דקדוקים שאינם בצורה נורמלית, אך לצורך פשטות השימוש נניח כי דקדוק הקלט הינו בצורה נורמלית.

קובץ הלקסיקון

מכיל את כל הטרמינלים (כלומר, את כל המילים ב- Σ).
יש להכניס כל מילה בפורמט הבא:

\w word ← כאן יש לכתוב את הטרמינל
\c nonterminal ← כאן יש לכתוב את הנונטרמינל שעובר לטרמינל

- בסוף שורה אין לשים סימן פיסוק.
- יש לכתוב לקובץ את המילים, כל מילה בצורה הנ"ל. אין הכרח להפריד בין מילה למילה ע"י שורה רווח.

כלומר, עבור כל חוק מן הצורה $A \to \alpha$ עבור $A \in V$ ו- $\alpha \in \Sigma$ נכתוב בקובץ הלקסיקון:

\w α
\c A

קובץ הדקדוק

מכיל את כל החוקים בשפה מן הצורה $A \to \alpha$ כאשר $\alpha \in V^*$. יש לשים לב לצורת הכתיבה:
כל חוק נכתב בשורה נפרדת ללא סימן פיסוק בסופו.
כל חוק ייכתב בצורה rule NonTerminal -> RuleBody.
המילה rule יכולה להיכתב כ- Rule או rule או RULE.

The terminal and nonterminal symbols in the rule have the following characteristics:

- Upper and lower case letters used in symbols are considered different. For example, `NOUN` is not the same as `Noun`, and neither is the same as `noun`.
- `The symbol X (capital letter x) may be used to stand for any terminal or nonterminal.`

`The symbol X can be useful for capturing generalities. Care must be taken, since it can be replaced by anything.`

- Index numbers are used to distinguish instances of a symbol that is used more than once in a rule. They are added to the end of a symbol following an underscore character (_).

For example the rule NP -> P NP should be written as :

Rule NP -> P NP_1

- The characters `(){}[]<>=:/` cannot be used in terminal or nonterminal symbols since they are used for special purposes in the grammar file. The character _ can be used *only* for attaching an index number to a symbol.
- By default, the left hand symbol of the first rule in the grammar file is the start symbol of the grammar.

The symbols on the right hand side of a phrase structure rule may be marked or grouped in various ways:

- Parentheses around an element of the expansion (right hand) part of a rule indicate that the element is optional. Parentheses may be placed around multiple elements. This makes an optional group of elements.
- A forward slash (/) is used to separate alternative elements of the expansion (right hand) part of a rule.
- Curly braces can be used for grouping alternative elements. For example the following says that an S consists of an NP followed by either a TVP or an IV:

```
Rule S -> NP {TVP / IV}
```

- Alternatives are taken to be as long as possible. Thus if the curly braces were omitted from the rule above, as in the rule below, the TVP would be treated as part of the alternative containing the NP. It would not be allowed before the IV:

```
Rule S -> NP TVP / IV
```

<u>קליטת קבצים וניתוח</u>

ראשית יש לתת לתוכנה לקלוט את הקבצים. שימו לב שכל סמל שיופיע בקובץ הדקדוק אך לא בקובץ הלקסיקון יזוהה כ- nonterminal.

קליטת קובץ הלקסיקון מתבצעת ע"י הפקודה: l l <filename>

קליטת קובץ הדקדוק מתבצעת ע"י הפקודה: l g <filename>

כעת אנו יכולים לבצע ניתוח על משפטים.

ע"י הפקודה parse <string> ניתן לבצע ניתוח למשפט בודד. הפלט יהיה תוצאת שייכות או אי שייכות לשפה, ובמידה והמשפט בשפה יודפסו לו כל עצי הניתוח האפשריים.

כמו כן ניתן לתת קובץ המכיל מספר משפטים לניתוח ע"י הפקודה

`file parse` *input-file [output-file]*

במידה וניתן קובץ פלט יודפסו אליו הניתוחים, במידה ולא, הם יופיעו על המסך.

סיום

כדי לצאת מהתוכנה יש להשתמש בפקודה exit.

דוגמאות

דוגמא מספר 1:

עבור השפה $L = \{a^n b^n \mid n \geq 0\}$ יש לבנות את הדקדוק הבא:

$$S \rightarrow A\ S\ B$$
$$S \rightarrow \varepsilon$$
$$A \rightarrow a$$
$$B \rightarrow b$$

קובץ הלקסיקון:

```
\w a
\c A

\w b
\c B
```

קובץ הדקדוק:

```
Rule S -> (A S_1 B)
```

דוגמא 2:

קובץ לקסיקון:

```
\w the
\c D

\w cat
\c N

\w hat
\c N

\w in
\c P
```

קובץ דקדוק:

```
Rule NP -> D N / NP_1 PP
Rule PP -> P  NP
```