TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

_____ ★★★ _____

KHOA: KHOA HỌC MÁY TÍNH

Môn Học : XLNNTN NÂNG CAO

Báo cáo đồ án cuối kỳ

XÂY DỰNG MỘT PARSER THEO MÔ HÌNH PCFG

(Probabilistic Context-Free Grammar)

GV Lý Thuyết: Nguyễn Tuấn Đăng

GV HDTH: Nguyễn Bích Vân

Sinh Viên thực hiện:

1. Lê Thành Đạt 13520199

2. Huỳnh Văn Tâm 13520735

TP.Hồ Chí Minh, tháng 12 năm 2016

LỜI CẢM ƠN

Trước tiên, chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Tiến sĩ Trần Tuấn Đăng -người đã tận tình chỉ bảo và hướng dẫn và góp ý cho chúng em trong quá trình thực hiện đồ án cuối kỳ.

Đồ án cuối kỳ "Xây Dựng Một Parser Theo Mô Hình PCFG" của môn Xử Lý Ngôn Ngữ Tự Nhiên là đồ án giúp cho chúng em hệ thống lại những kiến thức về ngữ pháp Tiếng Việt , văn phạm CNF và thuật toán CKY mà chúng em đã học.

Trong quá trình thực hiện đồ án , chúng em sẽ không tránh khỏi những thiếu sót về ngữ pháp trong văn phạm .Chúng em rất mong nhận được sự thông cảm và góp ý của quý thầy để chúng em có thể nắm bắt chắc hơn về kiến thức , giúp hoàn thiện và hiểu sâu về việc xử lý ngôn ngữ tự nhiên hơn.

Chúng em xin chân thành cảm ơn quý thầy!

MỤC LỤC

I. Các câu trong Treebank :

Tập hợp 20 câu trong Treebank :

1 . Nam học

2 . Nam làm bài

3 . Nam làm bài ở trường

4 . Nam làm bài của môn anh văn

5 . Nam thích đọc sách

6 . Nam đến nhà Lan

7. Lan đi chợ và nấu cơm

8 . Nam thường về quê

9 . Nhà của Nam ở quê

10 . Nam đang sống ở thành phố

11 . Nam học đại học ở thành phố

12 . Nam học ngành CNTT

13 . Nó sắp thi học kỳ

14 . Nó đang học bài

15 . Lan học với Nam

16 . Họ đang học ở trường

17 . Lan đã về nhà

18 . Lan và Nam đã học bài

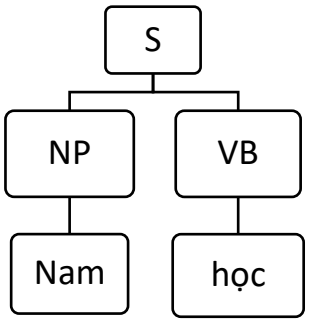
19 . Nam về nhà và đọc sách

20 . Nam thường nói chuyện với Lan

II. <u>Vẽ cây và xây dựng cấu trúc cú pháp cho từng câu trong Treebank</u> :

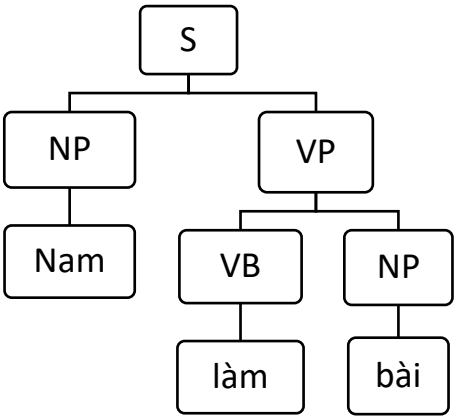
Phân tích câu theo văn phạm CNF (Chomsky Normal Form)

1 .Nam học.



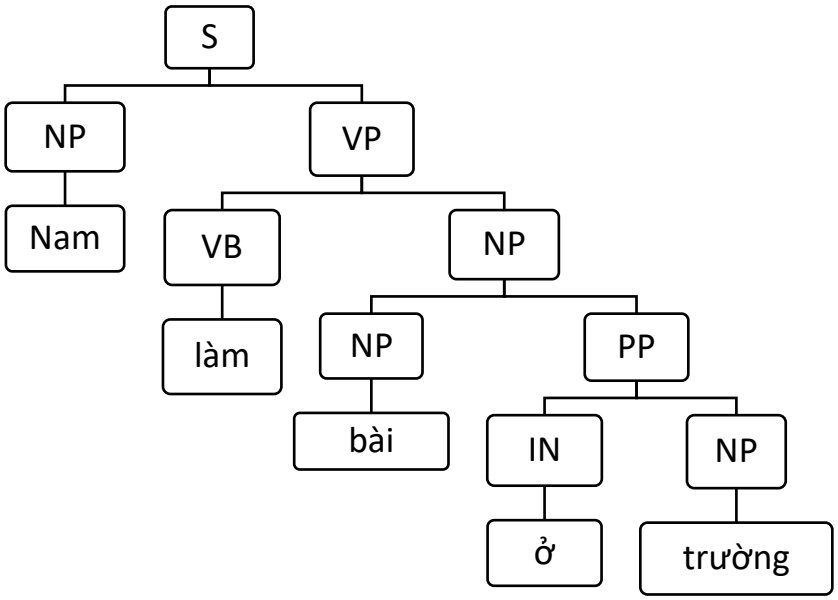
(S (NP Nam)(VB học))

2 .Nam làm bài.



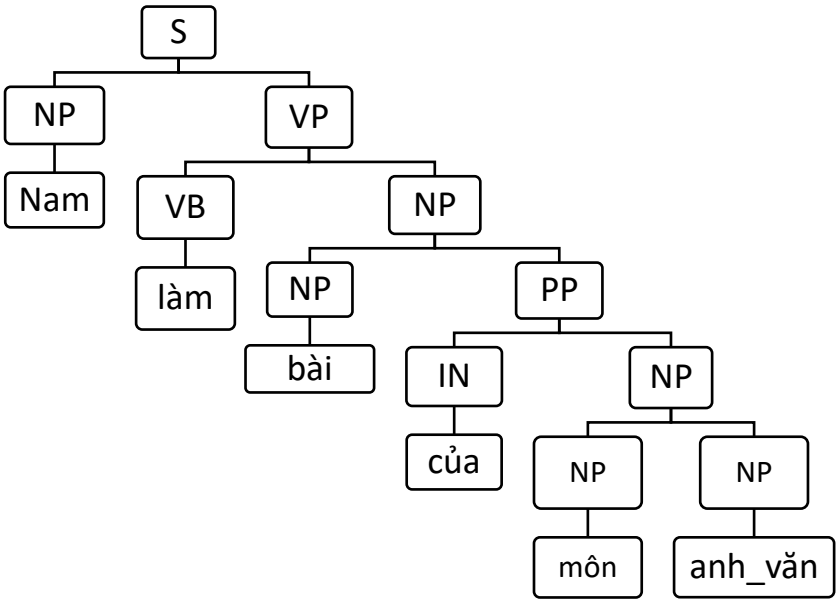
(S (NP Nam)(VP (VB làm)(NP bài)))

3 .Nam làm bài ở trường .



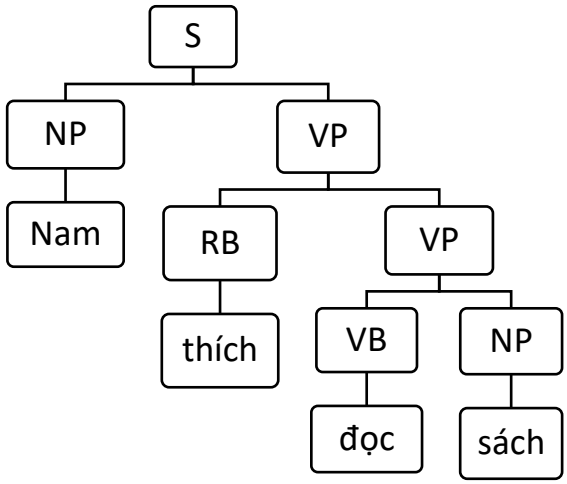
(S (NP Nam)(VP (VB làm)(NP (NP bài)(PP (IN ở)(NP trường)))))

4 .Nam làm bài của môn anh văn .



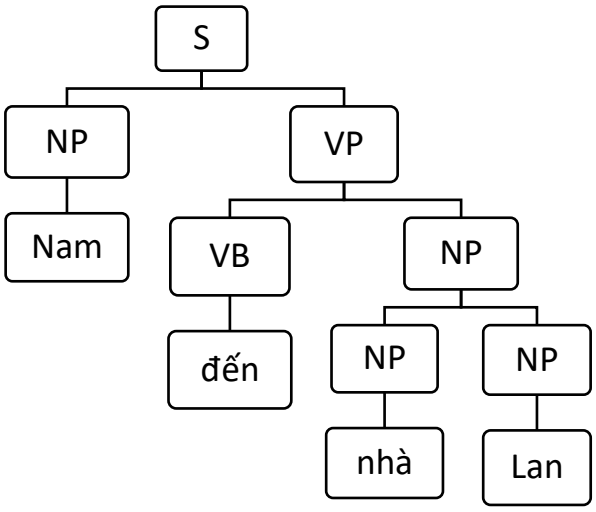
(S (NP Nam)(VP (VB làm)(NP (NP bài)(PP (IN của)(NP (NP môn)(NP anh_văn))))))

5 .Nam thích đọc sách.



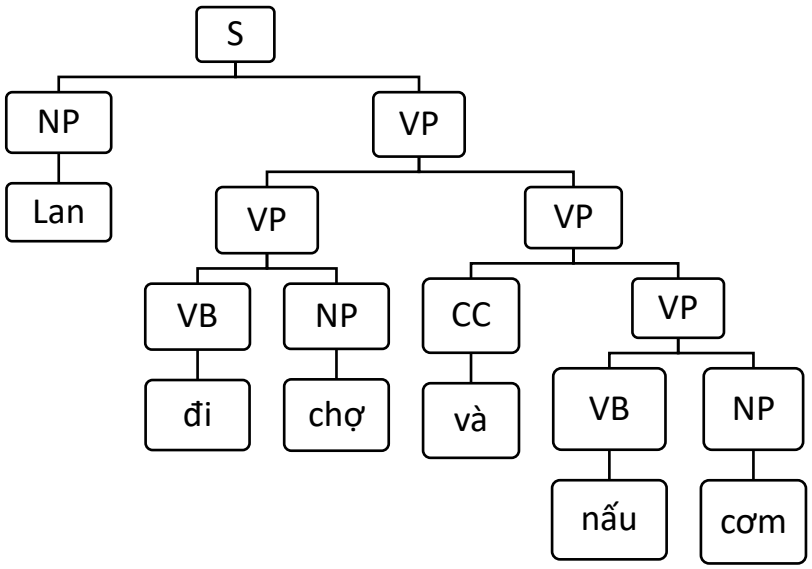
(S (NP Nam) (VP (RB thích)(VP (VB đọc)(NP sách))))

6 .Nam đến nhà Lan.



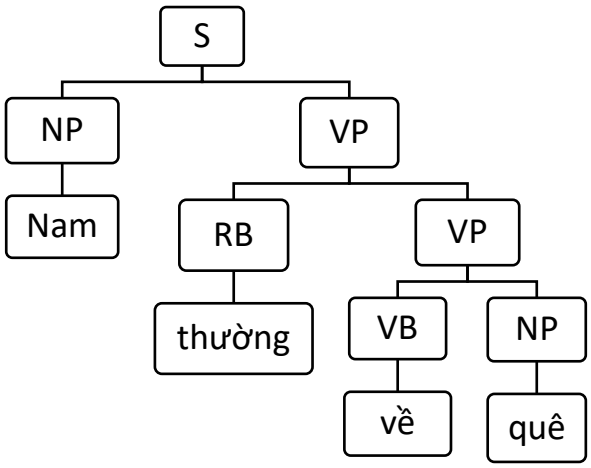
(S (NP Nam)(VP (VB đến)(NP (NP nhà)(NP Lan))))

7.Lan đi chợ và nấu cơm.



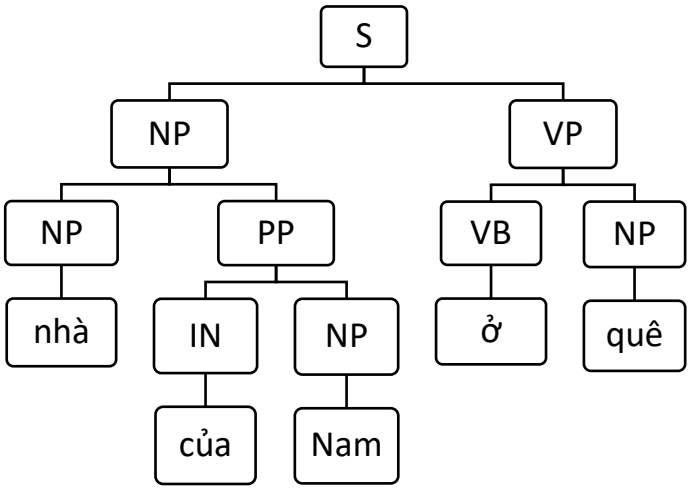
(S (NP Lan)(VP (VP (VB đi)(NP chợ))(VP (CC và)(VP (VB nấu)(NP cơm)))))

8 .Nam thường về quê.



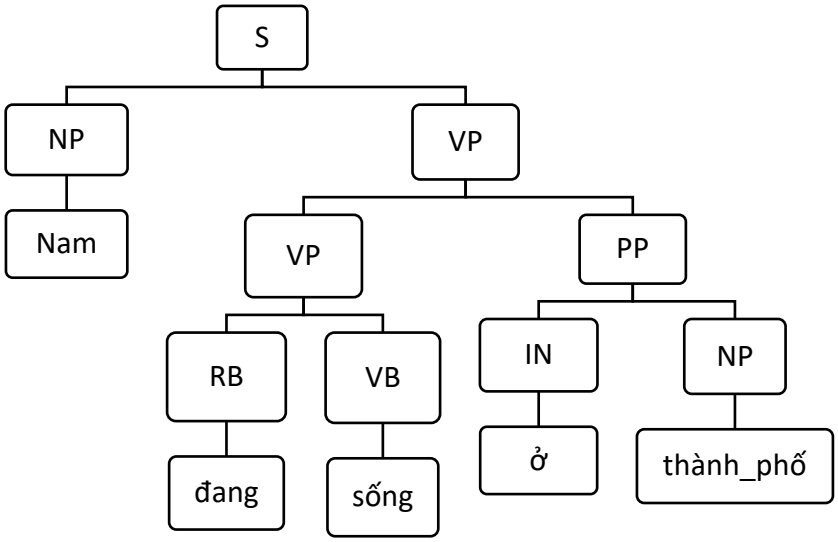
(S (NP Nam) (VP (RB thường)(VP (VB về)(NP quê))))

9 .Nhà của Nam ở quê.



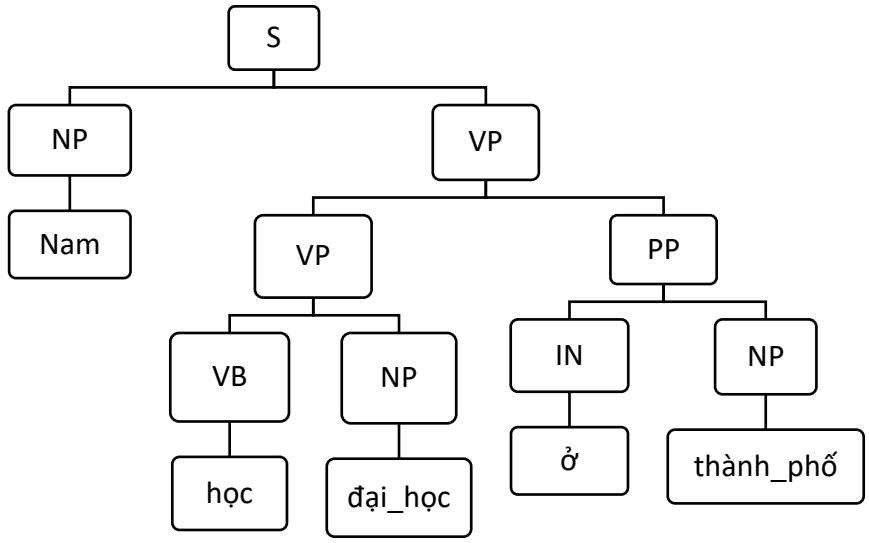
(S (NP (NP nhà)(PP (IN của)(NP Nam)))(VP (VB ở)(NP quê)))

10 .Nam đang sống ở thành phố.



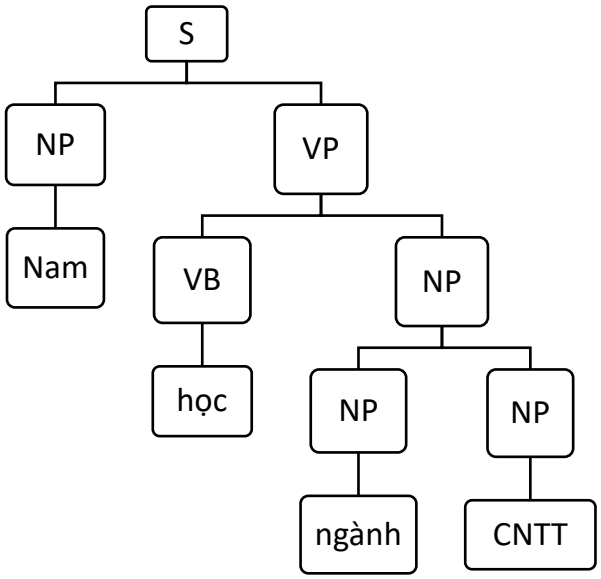
(S (NP Nam)(VP (VP (RB đang)(VB sống))(PP (IN ở)(NP thành_phố))))

11. Nam học đại học ở thành phố.



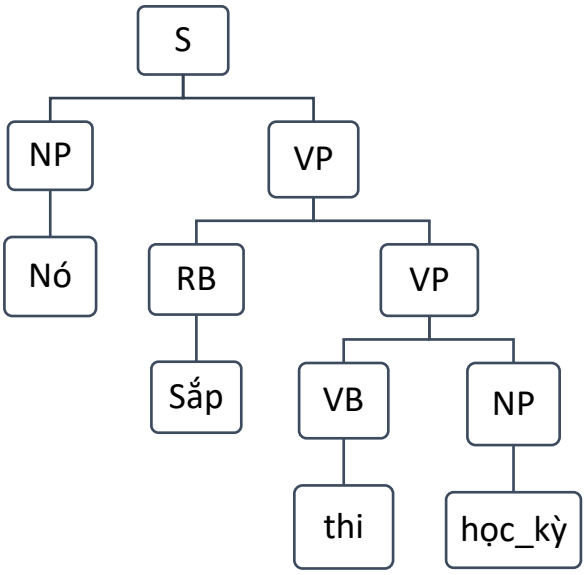
(S (NP Nam)(VP (VP (VB học)(NP đại_học))(PP (IN ở)(NP thành_phố))))

12. Nam học ngành CNTT.



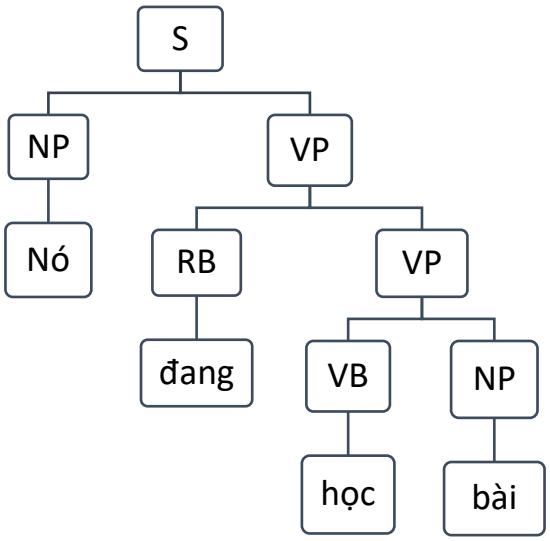
(S (NP Nam)(VP (VB học)(NP (NP ngành)(NP CNTT))))

13. Nó sắp thi học kỳ.



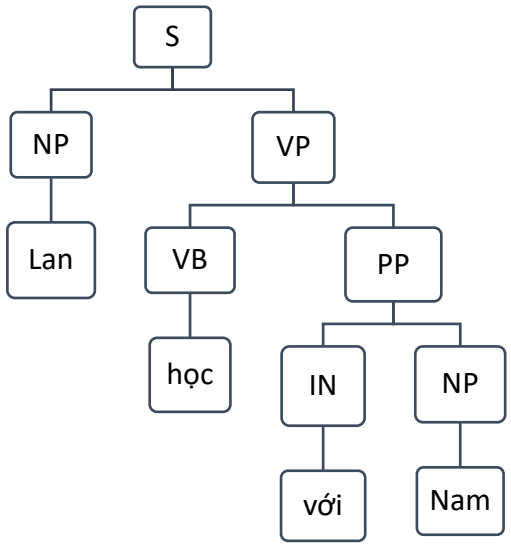
(S (NP nó)(VP (RB sắp)(VP (VB thi)(NP học_kỳ))))

14. Nó đang học bài.



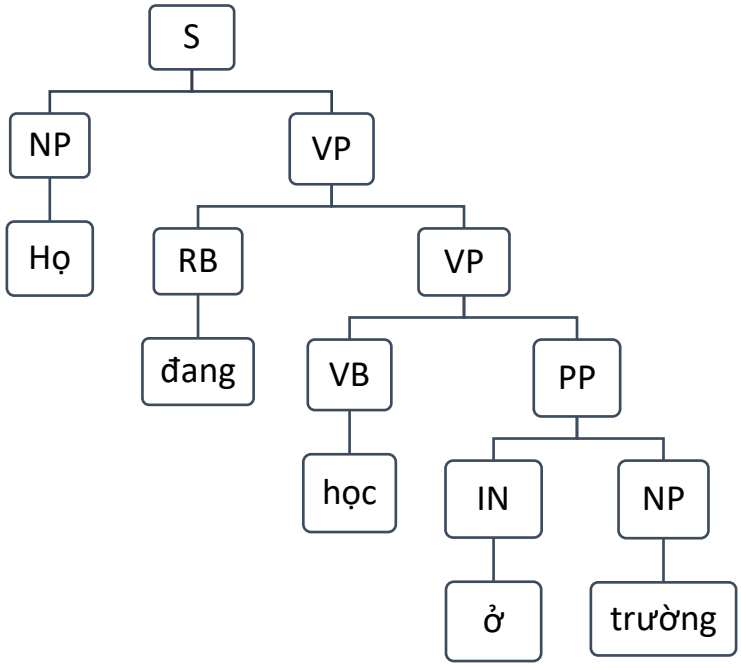
(S (NP nó)(VP (RB đang)(VP (VB học)(NP bài))))

15. Lan học với Nam.



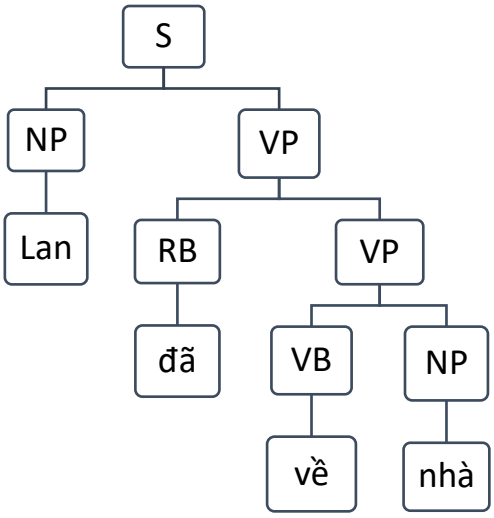
(S (NP Lan)(VP (VB học)(PP (IN với)(NP Nam))))

16. Họ đang học ở trường.



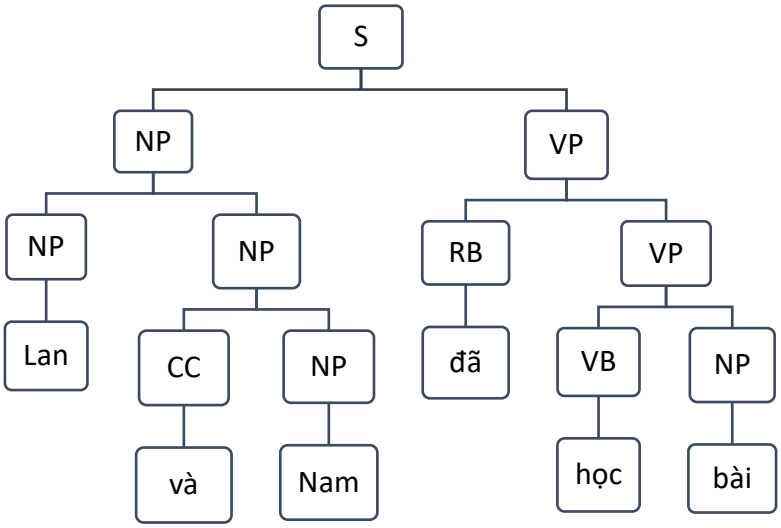
(S (NP họ)(VP (RB đang)(VP (VB học)(PP (IN ở)(NP trường)))))

17. Lan đã về nhà.



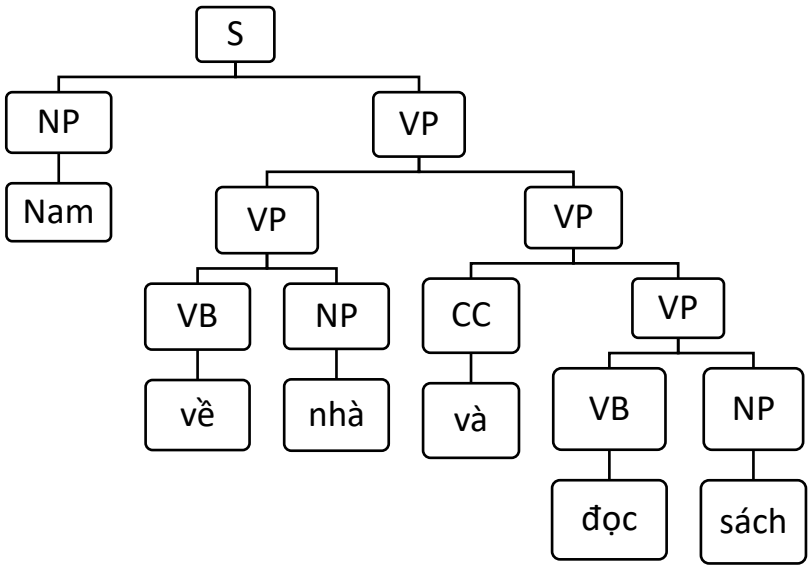
(S (NP Lan)(VP (RB đã)(VP (VB về)(NP nhà))))

18. Lan và Nam đã học bài.



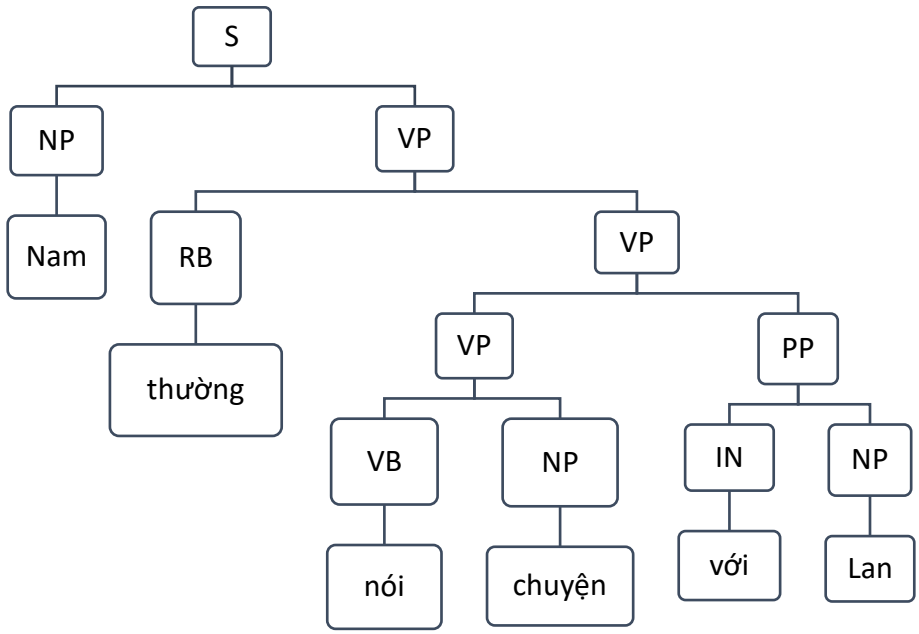
(S (NP (NP Lan)(NP (CC và)(NP Nam)))(VP (RB đã)(VP (VB học)(NP bài))))

19. Nam về nhà và đọc sách.



(S (NP Nam)(VP (VP (VB về)(NP nhà))(VP (CC và)(VP (VB đọc)(NP sách)))))

20. Nam thường nói chuyện với Lan.



(S (NP Nam)(VP (RB thường)(VP (VP (VB nói)(NP chuyện))(PP (IN với)(NP Lan)))))

TREEBANK :

(S (NP Nam)(VB học))

(S (NP Nam)(VP (VB làm)(NP bài)))

(S (NP Nam)(VP (VB làm)(NP (NP bài)(PP (IN ở)(NP trường)))))

(S (NP Nam)(VP (VB làm)(NP (NP bài)(PP (IN của)(NP (NP môn)(NP anh_văn))))))

(S (NP Nam) (VP (RB thích)(VP (VB đọc)(NP sách))))

(S (NP Nam)(VP (VB đến)(NP (NP nhà)(NP Lan))))

(S (NP Lan)(VP (VP (VB đi)(NP chợ))(VP (CC và)(VP (VB nấu)(NP cơm)))))

(S (NP Nam) (VP (RB thường)(VP (VB về)(NP quê))))

(S (NP (NP nhà)(PP (IN của)(NP Nam)))(VP (VB ở)(NP quê)))

(S (NP Nam)(VP (VP (RB đang)(VB sống))(PP (IN ở)(NP thành_phố))))

(S (NP Nam)(VP (VP (VB học)(NP đại_học))(PP (IN ở)(NP thành_phố))))

(S (NP Nam)(VP (VB học)(NP (NP ngành)(NP CNTT))))

(S (NP nó)(VP (RB sắp)(VP (VB thi)(NP học_kỳ))))

(S (NP nó)(VP (RB đang)(VP (VB học)(NP bài))))

(S (NP Lan)(VP (VB học)(PP (IN với)(NP Nam))))

(S (NP họ)(VP (RB đang)(VP (VB học)(PP (IN ở)(NP trường)))))

(S (NP Lan)(VP (RB đã)(VP (VB về)(NP nhà))))

(S (NP (NP Lan)(NP (CC và)(NP Nam)))(VP (RB đã)(VP (VB học)(NP bài))))

(S (NP Nam)(VP (VP (VB về)(NP nhà))(VP (CC và)(VP (VB đọc)(NP sách)))))

(S (NP Nam)(VP (RB thường)(VP (VP (VB nói)(NP chuyện))(PP (IN với)(NP Lan)))))

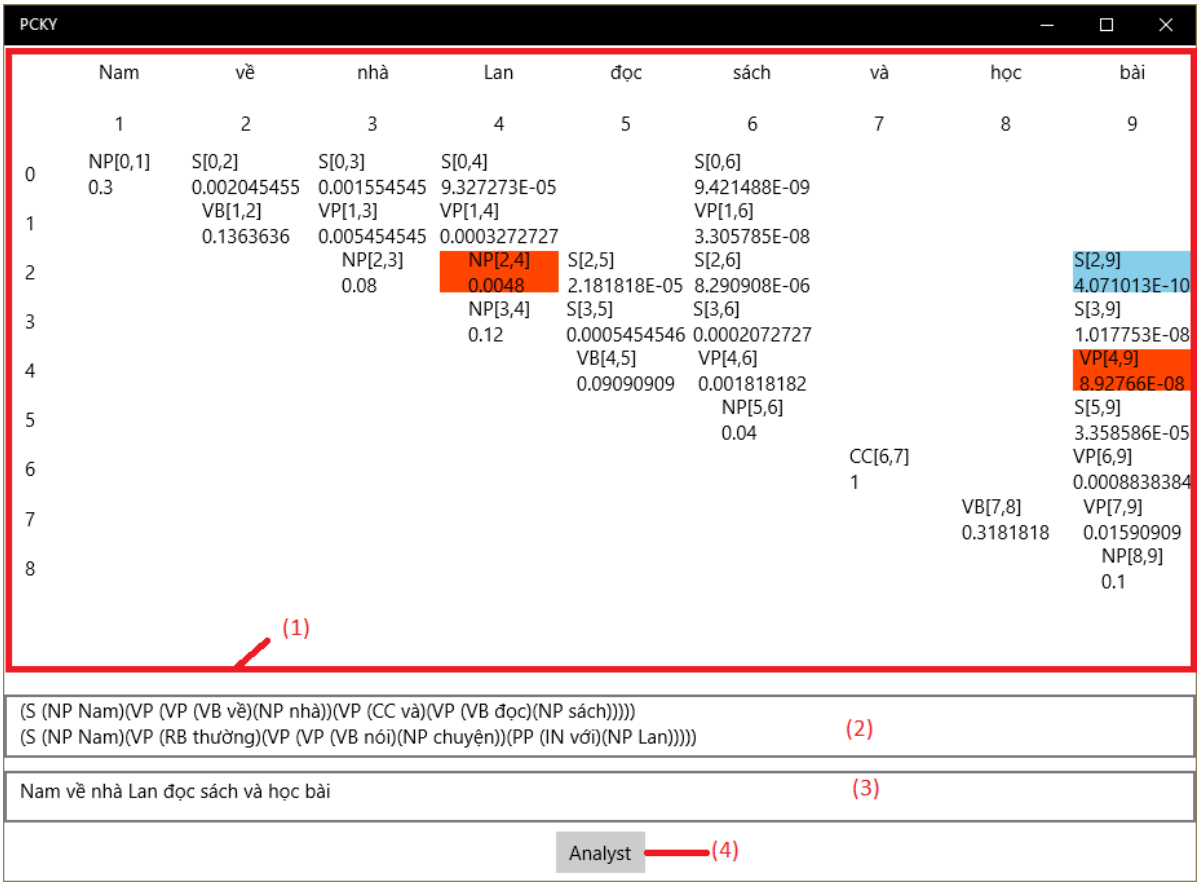
II. Một vài câu test mở rộng và kết quả test :

Bảng kết quả test các câu mở rộng trên Treebank đã xây dựng :

TT	Câu test	Kết quả test
1	Nam học CNTT	0.0009068181
2	Nam sắp thi học kỳ với Lan	7.996634E-09
3	Nam và Lan đang học bài với nó	2.099116E-09
4	Họ đang thi môn anh văn	6.397306E-09
5	Lan đã học bài và đọc sách ở nhà	1.675314E-12
6	Nam về nhà Lan đọc sách và học bài	4.626151E-13
7	họ thích học môn anh văn	1.492705E-08
8	Nam và Lan học bài thi học kỳ	8.587294E-10
9	Lan đã học bài ở trường với Nam	9.329405E-10
10	Nam đã về nhà với Lan đọc sách và nói chuyện với Lan	6.799165E-20
11	họ nói chuyện về bài thi học kỳ	8.260985E-14
12	Nam đã về quê với Lan	9.59596E-08
13	Nam thường nói chuyện với Lan về trường đại học	2.423222E-14
14	Lan học CNTT ở thành phố	6.045454E-07
15	Họ đọc sách và nói chuyện với Lan	1.211611E-13
16	Lan về với nó	8.636363E-06
17	Nam học đại học CNTT ở thành phố	1.511364E-08
18	Lan và Nam thích học ngành CNTT	1.679293E-09
19	Nam đang đi thành phố	1.919192E-05
20	Nam học bài anh văn	4.534091E-05

III. Giới thiệu giao diện chương trình và hướng dẫn sử dụng

NOTE : Ứng dụng PCKY của nhóm được làm trên nền tảng Universal Windows Platform (UWP) nên ứng dụng chỉ chạy trên win10 (từ build 10240 trở lên) .



(1) : Nơi thể hiện quá trình phân tích và tính xác suất cú pháp theo thuật toán PCKY.
(2) : Nơi để nhập TREEBANK
(3) : Nơi để nhập câu test
(4) : sau khi nhập Treebank và câu test , nhấp nút "Analyst" để chương trình thực thi việc tính xác suất theo PCKY

IV.Lời Cam Đoan

Nhóm chúng em xin cam đoan về nội dung đồ án cũng như souce code hoàn toàn do nhóm tự thiết kế và lập trình, hoàn toàn không sao chép nội dung cơ bản từ các đồ án khác , và sản phẩm tạo ra là do bản thân nhóm em tự mình nghiên cứu xây dựng nên.