**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

-----

**ĐỒ ÁN MÔN HỌC NGÔN NGỮ LẬP TRÌNH JAVA**

**PHẦN MỀM XỬ LÝ NGÔN NGỮ TỰ NHIÊN**



**GVHD: hUỲNH TUẤN ANH**

**nGUYỄN TẤN TOÀN**

**lớp: SE330.I21**

**svth: huỳnh quốc việt – 14521081**

**bùi quang trung - 14521014**

# NHẬN XÉT CỦA GIÁO VIÊN



# LỜI CẢM ƠN

Đầu tiên, chúng tôi xin gửi lời cảm ơn sâu sắc nhất tới thầy Huỳnh Tuấn Anh, người hướng dẫn khoa học, đã tận tình chỉ bảo, giúp đỡ và tạo điều kiện cho tôi hoàn thành đồ án này.

Chúng tôi xin chân thành cảm ơn các thầy cô trường Đại học Công nghệ Thông Tin, Đại học Quốc gia Hồ Chí Minh đã giảng dạy và truyền đạt kiến thức cho tôi.

Cuối cùng, chúng tôi xin cảm ơn những người thân và các bạn bè đã chia sẻ, giúp đỡ tôi hoàn thành đồ án này.

Mặc dù đã hết sức cố gắng với tất cả sự nỗ lực của bản thân, nhưng chắc đồ án vẫn còn những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý Thầy, Cô và bạn bè.

Tôi xin chân thành cảm ơn!

HCM, ngày 18 tháng 6 năm 2018

# LỜI MỞ ĐẦU

Từ những năm cuối của thế kỷ XX, những nền móng đầu tiên của nghành công nghệ thông tin (CNTT) đã ra đời. Từ đó đến nay, CNTT đã trở thành một trong nghành có tốc độ phát triển nhanh và ảnh hưởng sâu rộng nhất đến đời sống của con người. CNTT đã và đang thay đổi và cả thiện tất cả các mặt của cuộc sống, từ các nghành nghiên cứu khoa học cho đến hoạt động sản xuất kinh tế và cả chính trị, giải trí. Nhờ có CNTT mà cuộc sống ngày càng dễ dàng hơn, con người từ khắp mọi nơi trên thế giới có thể kết nối với nhau dễ hơn bao giờ hết. Nghành CNTT đã trở thành một trong những nghành kinh tế tạo ra nhiều giá trị nhất của các quốc gia.

Tuy nhiên, cùng với sự phát triển đó thì nhu cầu của con người và xã hội ngày một cao hơn. Con người không chỉ muốn điều khiển máy móc thực hiện mục tiêu của mình mà còn muốn các hệ thống ngày càng thông minh hơn, làm việc hiệu quả hơn. Những phần mềm, robot không chỉ là công cụ làm việc nữa mà trở thành một đồng nghiệp, một người bạn của con người hay thậm chí thay thế con người đưa ra quyết định giải quyết các vấn đề.

Đứng trước nhu cầu đó, nghành Khoa học máy tính nói chung và Trí tuệ nhân tạo nói riêng là một bộ phận đảm nhận trọng trách lớn nhất để tìm ra những giải pháp, những mô hình mới có thể hiện thực hóa những ý tưởng nhằm đáp ứng nhu cầu trên. Những sản phẩm tạo ra phải mang tính “thông minh”, có khả năng học hỏi, suy luận và “gần gũi” với con người hơn nữa.

Không chỉ có nhu cầu, hiện nay vấn đề ứng dụng CNTT vẫn còn phải được quan tâm giải quyết nhiều hơn. Nhà nước ta tuy đã chú trọng vào phát triển nguồn nhân lực CNTT nhiều và chất lượng hơn, nhưng nguồn nhân lực đó đa số vẫn sử dụng cho việc sản xuất kinh tế nhiều hơn cho việc nghiên cứu, giáo dục và thành tựu vẫn còn chưa đồng đều. Trong khi nhu cầu của học sinh hiện nay đang cần có một công cụ giúp hỗ trợ xử lý ngôn ngữ tự nhiên vào máy tính.

Do đó, chúng em đã xây dựng chương trình phân tích cú pháp câu trong bộ ngữ liệu huấn luyện để xử lý ngôn ngữ tự nhiên. Với mục tiêu là giúp học sinh và giáo viên có thể phân tích ngôn ngữ tự nhiên trên máy tính.

# MỤC LỤC

[NHẬN XÉT CỦA GIÁO VIÊN 3](#_Toc517095000)

[LỜI CẢM ƠN 4](#_Toc517095001)

[LỜI MỞ ĐẦU 5](#_Toc517095002)

[MỤC LỤC 7](#_Toc517095003)

[Chương 1 HIỆN TRẠNG VÀ YÊU CẦU THỰC TẾ 9](#_Toc517095004)

[1.1 Hiện trạng vấn đề 9](#_Toc517095005)

[1.1.1 Vấn đề 9](#_Toc517095006)

[1.1.2 Phương pháp giải quyết 12](#_Toc517095007)

[1.2 Hiện trạng cơ sở vật chất và con người 12](#_Toc517095008)

[1.2.1 Cơ sở vật chất 12](#_Toc517095009)

[1.2.2 Con người 12](#_Toc517095010)

[1.3 Yêu cầu sơ bộ về phần mềm 13](#_Toc517095011)

[1.3.1 Tin học 13](#_Toc517095012)

[1.3.2 Con người 13](#_Toc517095013)

[1.4 Yêu cầu đồ án 13](#_Toc517095014)

[Chương 2 THIẾT KẾ 15](#_Toc517095015)

[2.1 Thiết kế giao diện chính 15](#_Toc517095016)

[2.1.1 Giao diện ban đầu 15](#_Toc517095017)

[2.1.2 Giao diện chạy chương trình 16](#_Toc517095018)

[2.2 Các nút lệnh và thao tác 17](#_Toc517095019)

[2.2.1 Grammar have probability 17](#_Toc517095020)

[2.2.2 Input grammar 18](#_Toc517095021)

[2.2.3 Nút Probability 18](#_Toc517095022)

[2.2.4 Sentence Train 18](#_Toc517095023)

[2.2.5 Tree 19](#_Toc517095024)

[2.2.6 ActionP-CKY 20](#_Toc517095025)

[2.2.7 ActionP – CKY 20](#_Toc517095026)

[Chương 3 CÀI ĐẶT PHẦN MỀM 21](#_Toc517095027)

[3.1 Tổng quan về công nghệ 21](#_Toc517095028)

[3.2 Cài đặt 22](#_Toc517095029)

[3.2.1 Hình vẽ cây cú pháp 22](#_Toc517095030)

[Chương 4 KẾT QUẢ PHÂN TÍCH CÚ PHÁP BẰNG THUẬT TOÁN PCKY 33](#_Toc517095031)

[4.1 Ví dụ 1: Nam học bài ở trường 33](#_Toc517095032)

[4.2 Ví dụ 2: Nam rất thích đọc sách ở thư viện. 34](#_Toc517095033)

[Chương 5 tỔNG KẾT 35](#_Toc517095034)

[5.1 Kết quả đạt được 35](#_Toc517095035)

[5.2 Hạn chế 35](#_Toc517095036)

[5.3 Hướng phát triển 35](#_Toc517095037)

[TÀI LIỆU THAM KHẢO 36](#_Toc517095038)

# HIỆN TRẠNG VÀ YÊU CẦU THỰC TẾ

## Hiện trạng vấn đề

### Vấn đề

* Xử lý ngôn ngữ tự nhiên (*natural language processing* - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.
* Phân tích hình thái - Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong [tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong [tiếng Việt](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Vi%E1%BB%87t), dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ. Cùng với các ngôn ngữ như [tiếng Trung](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Trung_Qu%E1%BB%91c), [tiếng Hàn](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Tri%E1%BB%81u_Ti%C3%AAn),[tiếng Nhật](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Nh%E1%BA%ADt), [phân tách từ](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%A1ch_t%E1%BB%AB&action=edit&redlink=1) trong tiếng Việt là một công việc không hề đơn giản.
* Phân tích cú pháp - Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các luật văn phạm.
* Phân tích ngữ nghĩa - Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp.
* Tích hợp văn bản - Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.
* Phân tích thực nghĩa - Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.
* Thời đại hiện hay máy tính ngày càng thông minh, có thể giải quyết các vấn đề con người đặt ra. Trong đó mảng xử lý ngôn ngữ tự nhiên ngày càng được quan tâm.
* [Nhận dạng chữ viết](https://vi.wikipedia.org/w/index.php?title=Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt&action=edit&redlink=1): Có hai kiểu nhận dạng, thứ nhất là nhận dạng chữ in, ví dụ nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc của [Microsoft Word](https://vi.wikipedia.org/wiki/Microsoft_Word) chẳng hạn. Phức tạp hơn là [nhận dạng chữ viết tay](https://vi.wikipedia.org/w/index.php?title=Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt_tay&action=edit&redlink=1), có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng và thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).
* [Nhận dạng tiếng nói](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i): Nhận dạng tiếng nói rồi chuyển chúng thành văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.
* [Tổng hợp tiếng nói](https://vi.wikipedia.org/wiki/T%E1%BB%95ng_h%E1%BB%A3p_gi%E1%BB%8Dng_n%C3%B3i): Từ một văn bản tự động tổng hợp thành tiếng nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang [web](https://vi.wikipedia.org/wiki/World_Wide_Web), nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, tổng hợp tiếng nói là sự trợ giúp tốt cho [người khiếm thị](https://vi.wikipedia.org/wiki/Ng%C6%B0%E1%BB%9Di_khi%E1%BA%BFm_th%E1%BB%8B), nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa [robot](https://vi.wikipedia.org/wiki/Robot) với người.
* [Dịch tự động](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng) (*machine translate*): Như tên gọi đây là chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Một phần mềm điển hình về tiếng Việt của chương trình này là [Evtrans](https://vi.wikipedia.org/w/index.php?title=Evtrans&action=edit&redlink=1) của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web vdict.com mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Tháng 10 năm 2008 có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và [Google](https://vi.wikipedia.org/wiki/Google), một thời gian sau đó Xalo\_vn cũng đưa ra dịch vụ tương tự.
* [Tìm kiếm thông tin](https://vi.wikipedia.org/wiki/T%C3%ACm_ki%E1%BA%BFm_th%C3%B4ng_tin) (*information retrieval*): Đặt câu hỏi và chương trình tự tìm ra nội dung phù hợp nhất. Thông tin ngày càng đầy lên theo cấp số nhân, đặc biệt với sự trợ giúp của [internet](https://vi.wikipedia.org/wiki/Internet) việc tiếp cận thông tin trở lên dễ dàng hơn bao giờ hết. Việc khó khăn lúc này là tìm đúng nhất thông tin mình cần giữa bề bộn tri thức và đặc biệt thông tin đó phải đáng tin cậy. Các [máy tìm kiếm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) dựa trên giao diện web như Google hay [Yahoo](https://vi.wikipedia.org/wiki/Yahoo!) hiện nay chỉ phân tích nội dung rất đơn giản dựa trên tần suất của từ khoá và thứ hạng của trang và một số tiêu chí đánh giá khác để đưa ra kết luận, kết quả là rất nhiều tìm kiếm không nhận được câu trả lời phù hợp, thậm chí bị dẫn tới một liên kết không liên quan gì do thủ thuật đánh lừa của các trang web nhằm giới thiệu sản phẩm (có tên tiếng Anh là [SEO](https://vi.wikipedia.org/wiki/T%E1%BB%91i_%C6%B0u_h%C3%B3a_c%C3%B4ng_c%E1%BB%A5_t%C3%ACm_ki%E1%BA%BFm) viết tắt của từ *search engine optimization*). Thực tế cho đến bây giờ chưa có máy tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người trừ trang www.ask.com được đánh giá là "hiểu" được những câu hỏi có cấu trúc ở dạng đơn giản nhất. Mới đây cộng đồng mạng đang xôn xao về trang [Wolfram Alpha](https://vi.wikipedia.org/wiki/Wolfram_Alpha), được hứa hẹn là có khả năng hiểu ngôn ngữ tự nhiên của con người và đưa ra câu trả lời chính xác. Lĩnh vực này hứa hẹn tạo ra bước nhảy trong cách thức tiếp nhận tri thức của cả cộng đồng.
* [Tóm tắt văn bản](https://vi.wikipedia.org/w/index.php?title=T%C3%B3m_t%E1%BA%AFt_v%C4%83n_b%E1%BA%A3n&action=edit&redlink=1): Từ một văn bản dài tóm tắt thành một văn bản ngắn hơn theo mong muốn nhưng vẫn chứa những nội dung thiết yếu nhất.
* [Khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u) (*data mining*) và [phát hiện tri thức](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A1t_hi%E1%BB%87n_tri_th%E1%BB%A9c&action=edit&redlink=1): Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình [học tập](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_t%E1%BA%ADp), khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển. Ở mức độ đơn giản khi kết hợp với [máy tìm kiếm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) nó cho phép đặt câu hỏi để từ đó công cụ tự tìm ra câu trả lời dựa trên các thông tin trên web mặc cho việc trước đó có câu trả lời lưu trên web hay không (giống như trang Yahoo! hỏi và đáp, nơi chuyên đặt các câu hỏi để người khác trả lời), nói một cách nôm na là nó đã biết xử lý dữ liệu để trả lời câu hỏi của người sử dụng, thay vì máy móc đáp trả những gì chỉ có sẵn trong bộ nhớ.

### Phương pháp giải quyết

* CKY ( Coke- Kasami – Younger) là một thuật toán cải tiến của thuật toán phân tích cú pháp Bottom-Up (Button-Up Parsing là một chiến lượt phân tích tích cú pháp bắt đầu từ các từ trong các chuỗi đầu vào và xây dựng các thành tố cú pháp).
* CKY có thể tránh được những cách phân tích cú pháp không hợp lý so với thuật toán Buttom-Up thông thường. Do CKY sử dụng một hình thức văn phạm đặc biệt được gọi là Chomsky Normal Form (CNF). Các giải pháp trung gian của thuật  được lưu trữ và chỉ triển khai những giải pháp trung gian nào có khả năng đóng góp vào việc phân tích đầy đủ cấu trúc cú pháp câu.

## Hiện trạng cơ sở vật chất và con người

### Cơ sở vật chất

Máy tính ngày càng giúp con người giải quyết những vấn đề mà con người không thể giải quyết được và muốn máy tính hiểu ngôn ngữ của con người thì chúng tôi đã tạo ra phần mềm giúp xử lý ngôn ngữ tự nhiên trên máy tính, giúp máy tính có thể hiểu ngôn ngữ tự nhiên.

### Con người

Thời đại hiện nay, công nghệ thông tin ngày càng phát triển và sản phẩm của sự phát triển đó là những chiếc máy tính thông minh được ra đời và ngày càng hoàn thiện hơn. Máy tính muốn được sử dụng rộng rãi thì ngôn ngữ máy tính phải thật gần gũi, giúp người dùng dễ dàng thao tác và sử dụng thì cần phải có các công cụ hỗ trợ để xử lý ngôn ngữ tự nhiên trên máy tính. Giúp máy tính hiểu được ngôn ngữ tự nhiên cảu người dùng và thực hiện yêu cầu người dùng.

## Yêu cầu sơ bộ về phần mềm

### Tin học

* Phần mềm thích hợp trên gần như tất cả các máy tính từ cấu hình trung bình đến cấu hình cao.
* Phần cứng không đòi hỏi quá cao.
* Cấu hình mẫu:
* Processor: Intel (R) Core (TM) i5-6200U CPU @ 2.30GHz 2.40GHz
* RAM: 4GB
* Operating System: Windows 10 Home,
* VGA: AMD Radeon R7 M360 (Total memory: 6088MB)
* Công cụ:
* Microsoft Visual Studio Community 2015 ( Ver 14.0)
* Netbean IDE 8.2
* Ngôn ngữ lập trình:
* Java
* Prolog

### Con người

* Ngôn ngữ chính: Tiếng Anh
* Giao diện: Dễ nhìn, dễ sử dụng.

## Yêu cầu đồ án

Đồ án yêu cầu xây dựng một parser dựa trên mô hình PCFG để phân tích cú pháp tối thiểu 10 câu mở rộng từ các câu trong ngữ liệu huấn luyện.

Ngữ liệu huấn luyện:

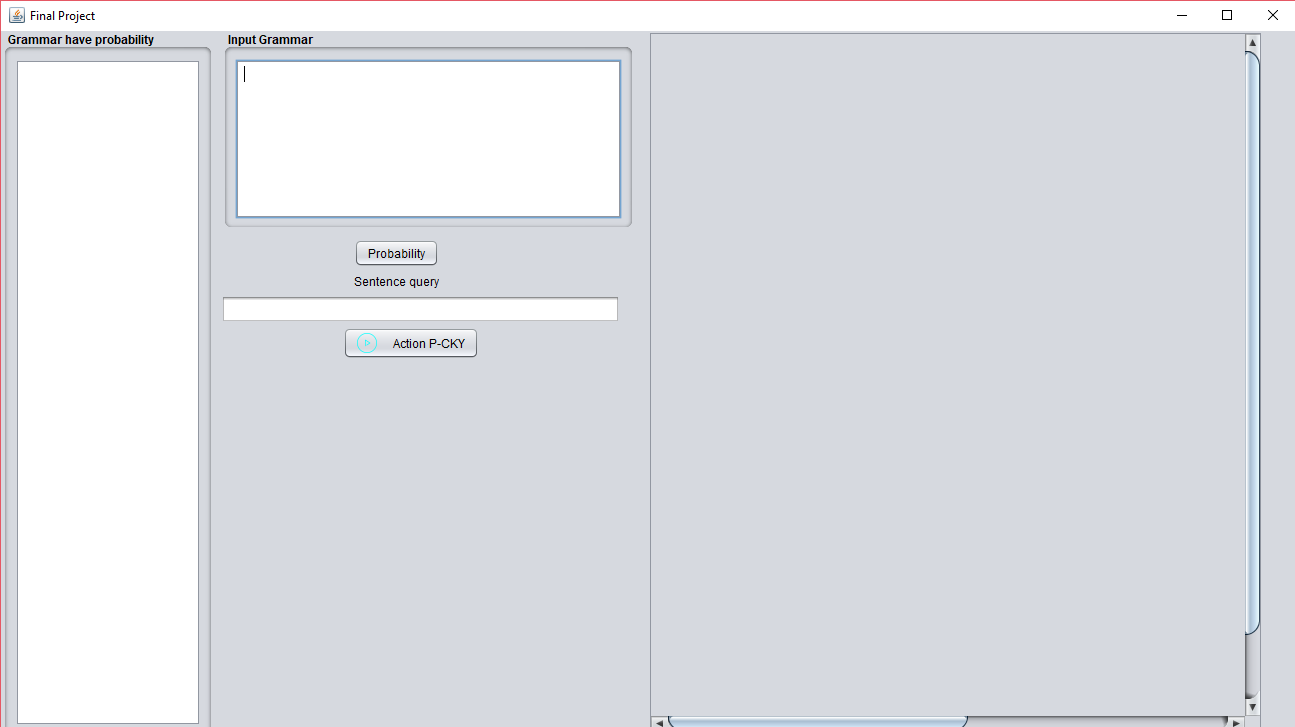
1. Nam học bài.
2. Nam đọc sách ở trường.
3. Nam ở trường.
4. Nam thích đọc sách.
5. Nam rất thích đến thư viện.
6. Nam và Lan đều thích đọc sách.
7. Nam vừa gặp Lan ở trường.
8. Nam cho Lan mượn một cuốn sách rất hay.
9. Lan hay mượn sách của Nam.
10. Nam có nhiều sách mới.

* Chương trình nhận ngữ liệu là cây cú pháp của các câu huấn luyện theo CNF, biểu diễn dưới dạng danh sách. Ví dụ: (S (NP ...) (VP ...))
* Chương trình rút ra các quy tắc cú pháp CNF từ ngữ liệu huấn luyện và tính xác suất cho các quy tắc.
* Chương trình sử dụng thuật toán PCKY để phân tích cú pháp cho các câu thử nghiệm và xuất kết quả.

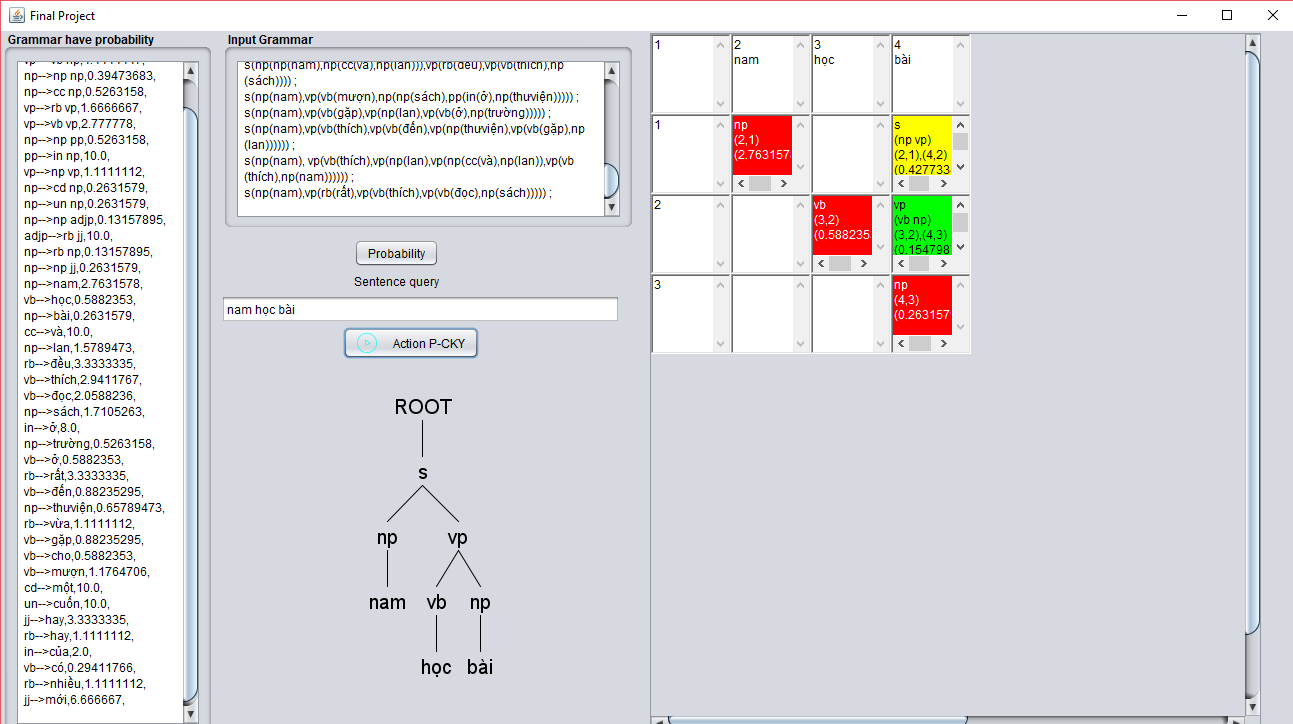
# THIẾT KẾ

## Thiết kế giao diện chính

### Giao diện ban đầu

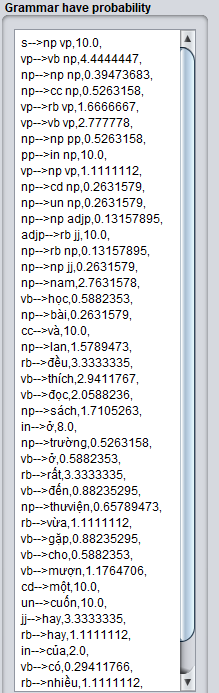


### Giao diện chạy chương trình



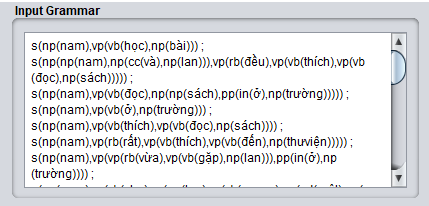
## Các nút lệnh và thao tác

### Grammar have probability



Grammar have probability: Xác suất của các từ trong bộ huấn luyện.

### Input grammar



Input grammar: Bộ huấn luyện của chương trình.

### Nút Probability



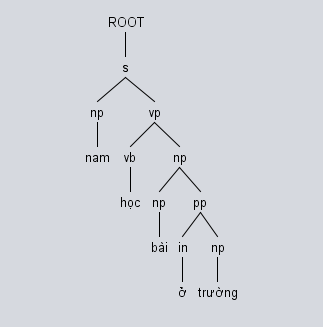
Nút Probability: Thực hiện tính các xác suất của các từ trong bộ huấn luyện.

### Sentence Train



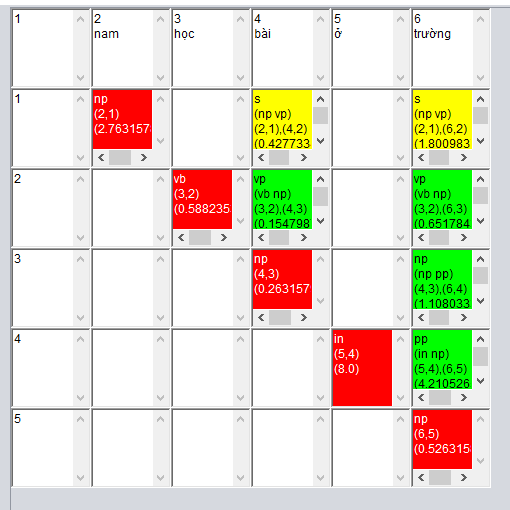
Sentence Train: Thực hiện vẽ cây CNF và vẽ cây

### Tree



Tree: Cây cú pháp của câu thử nghiệm.

### ActionP-CKY



ActionP-CKY: Thực hiện văn phạm P-CKY

### ActionP – CKY



ActionP – CKY: chạy P-CKY

# CÀI ĐẶT PHẦN MỀM

## Tổng quan về công nghệ

Java là một [ngôn ngữ lập trình](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh) [hướng đối tượng](https://vi.wikipedia.org/wiki/L%E1%BA%ADp_tr%C3%ACnh_h%C6%B0%E1%BB%9Bng_%C4%91%E1%BB%91i_t%C6%B0%E1%BB%A3ng) (OOP) và dựa trên các lớp (class). Khác với phần lớn ngôn ngữ lập trình thông thường, thay vì [biên dịch](https://vi.wikipedia.org/wiki/Tr%C3%ACnh_bi%C3%AAn_d%E1%BB%8Bch) [mã nguồn](https://vi.wikipedia.org/wiki/M%C3%A3_ngu%E1%BB%93n) thành [mã máy](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_m%C3%A1y) hoặc [thông dịch](https://vi.wikipedia.org/wiki/Tr%C3%ACnh_th%C3%B4ng_d%E1%BB%8Bch) mã nguồn khi chạy, Java được thiết kế để biên dịch mã nguồn thành [bytecode](https://vi.wikipedia.org/wiki/Bytecode), bytecode sau đó sẽ được môi trường thực thi (runtime environment) chạy.

Trước đây, Java chạy chậm hơn những ngôn ngữ dịch thẳng ra mã máy như C và [C++](https://vi.wikipedia.org/wiki/C%2B%2B), nhưng sau này nhờ công nghệ "biên dịch tại chỗ" - Just in time compilation, khoảng cách này đã được thu hẹp, và trong một số trường hợp đặc biệt Java có thể chạy nhanh hơn. Java chạy nhanh hơn những ngôn ngữ thông dịch như [Python](https://vi.wikipedia.org/wiki/Python), [Perl](https://vi.wikipedia.org/wiki/Perl), [PHP](https://vi.wikipedia.org/wiki/PHP) gấp nhiều lần. Java chạy tương đương so với [C#](https://vi.wikipedia.org/wiki/C_th%C4%83ng), một ngôn ngữ khá tương đồng về mặt cú pháp và quá trình dịch/chạy

[Cú pháp](https://vi.wikipedia.org/wiki/C%C3%BA_ph%C3%A1p_h%E1%BB%8Dc) Java được vay mượn nhiều từ [C](https://vi.wikipedia.org/wiki/C_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)) & [C++](https://vi.wikipedia.org/wiki/C%2B%2B) nhưng có cú pháp hướng đối tượng đơn giản hơn và ít tính năng xử lý cấp thấp hơn. Do đó việc viết một chương trình bằng Java dễ hơn, đơn giản hơn, đỡ tốn công sửa lỗi hơn.

Trong Java, hiện tượng rò rỉ bộ nhớ hầu như không xảy ra do bộ nhớ được quản lý bởi Java Virtual Machine (JVM) bằng cách tự động "dọn dẹp rác". Người lập trình không phải quan tâm đến việc cấp phát và xóa bộ nhớ như C, C++. Tuy nhiên khi sử dụng những tài nguyên mạng, file IO, database (nằm ngoài kiểm soát của JVM) mà người lập trình không đóng (close) các streams thì rò rỉ dữ liệu vẫn có thể xảy ra.

Java được khởi đầu bởi [James Gosling](https://vi.wikipedia.org/wiki/James_Gosling) và bạn đồng nghiệp ở [Sun Microsystems](https://vi.wikipedia.org/wiki/Sun_Microsystems) năm [1991](https://vi.wikipedia.org/wiki/1991). Ban đầu ngôn ngữ này được gọi là Oak (có nghĩa là [cây sồi](https://vi.wikipedia.org/w/index.php?title=C%C3%A2y_s%E1%BB%93i&action=edit&redlink=1); do bên ngoài cơ quan của ông Gosling có trồng nhiều loại cây này), họ dự định ngôn ngữ đó thay cho [C++](https://vi.wikipedia.org/wiki/C%2B%2B), nhưng các tính năng giống [Objective C](https://vi.wikipedia.org/w/index.php?title=Objective_C&action=edit&redlink=1). Không nên lẫn lộn Java với [JavaScript](https://vi.wikipedia.org/wiki/JavaScript), hai ngôn ngữ đó chỉ giống tên và loại [cú pháp](https://vi.wikipedia.org/wiki/C%C3%BA_ph%C3%A1p_h%E1%BB%8Dc)như [C](https://vi.wikipedia.org/wiki/C_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)). Công ty Sun Microsystems đang giữ bản quyền và phát triển Java thường xuyên. Tháng 04/2011, công ty Sun Microsystems tiếp tục cho ra bản JDK 1.6.24.

Java được tạo ra với tiêu chí "Viết (code) một lần, thực thi khắp nơi" ("Write Once, Run Anywhere" (WORA)). Chương trình phần mềm viết bằng Java có thể chạy trên mọi nền tảng (platform) khác nhau thông qua một môi trường thực thi với điều kiện có môi trường thực thi thích hợp hỗ trợ nền tảng đó. Môi trường thực thi của Sun Microsystems hiện hỗ trợ [Sun Solaris](https://vi.wikipedia.org/w/index.php?title=Sun_Solaris&action=edit&redlink=1), [Linux](https://vi.wikipedia.org/wiki/Linux), [Mac OS](https://vi.wikipedia.org/wiki/Mac_OS), [FreeBSD](https://vi.wikipedia.org/wiki/FreeBSD) & [Windows](https://vi.wikipedia.org/wiki/Microsoft_Windows). Ngoài ra, một số công ty, tổ chức cũng như cá nhân khác cũng phát triển môi trường thực thi Java cho những hệ điều hành khác như BEA, IBM, HP.... Trong đó đáng nói đến nhất là IBM Java Platform hỗ trợ Windows, Linux, [AIX](https://vi.wikipedia.org/w/index.php?title=AIX&action=edit&redlink=1) & [z/OS](https://vi.wikipedia.org/w/index.php?title=Z/OS&action=edit&redlink=1).

Những chi tiết về ngôn ngữ, [máy ảo](https://vi.wikipedia.org/w/index.php?title=M%C3%A1y_%E1%BA%A3o_Java&action=edit&redlink=1) và [API](https://vi.wikipedia.org/wiki/Giao_di%E1%BB%87n_l%E1%BA%ADp_tr%C3%ACnh_%E1%BB%A9ng_d%E1%BB%A5ng) của Java được giữ bởi [Cộng đồng Java](https://vi.wikipedia.org/w/index.php?title=C%E1%BB%99ng_%C4%91%E1%BB%93ng_Java&action=edit&redlink=1) (do [Sun](https://vi.wikipedia.org/wiki/Sun) quản lý). Java được tạo ra vào năm [1991](https://vi.wikipedia.org/wiki/1991) do một số kỹ sư ở [Sun](https://vi.wikipedia.org/wiki/Sun_Microsystems), bao gồm ông James Gosling, một phần của *Dự án Xanh* (Green Project). Java được phát hành vào năm [1994](https://vi.wikipedia.org/wiki/1994), rồi nó trở nên nổi tiếng khi [Netscape](https://vi.wikipedia.org/wiki/Netscape) tuyên bố tại hội thảo SunWorld năm [1995](https://vi.wikipedia.org/wiki/1995) là trình duyệt [Navigator](https://vi.wikipedia.org/wiki/Netscape_Navigator) của họ sẽ hỗ trợ Java. Về sau Java được hỗ trợ trên hầu hết các trình duyệt như [Internet Explorer](https://vi.wikipedia.org/wiki/Internet_Explorer) ([Microsoft](https://vi.wikipedia.org/wiki/Microsoft)), [Firefox](https://vi.wikipedia.org/wiki/Mozilla_Firefox) ([Mozilla](https://vi.wikipedia.org/wiki/Mozilla)), [Safari](https://vi.wikipedia.org/wiki/Safari) ([Apple](https://vi.wikipedia.org/wiki/Apple_Inc.))...

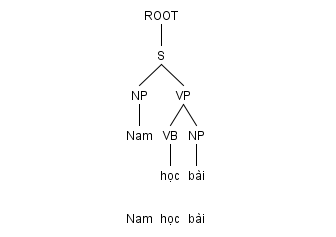
Java được sử dụng chủ yếu trên môi trường NetBeans và Oracle. Sau khi [Oracle](https://vi.wikipedia.org/wiki/Oracle) mua lại [công ty](https://vi.wikipedia.org/wiki/C%C3%B4ng_ty) [Sun Microsystems](https://vi.wikipedia.org/wiki/Sun_Microsystems) năm [2009](https://vi.wikipedia.org/wiki/2009)-[2010](https://vi.wikipedia.org/wiki/2010), [Oracle](https://vi.wikipedia.org/wiki/Oracle) đã mô tả họ là "*người quản lý công nghệ Java với cam kết không ngừng để bồi dưỡng một cộng đồng tham gia và minh bạch*".

## Cài đặt

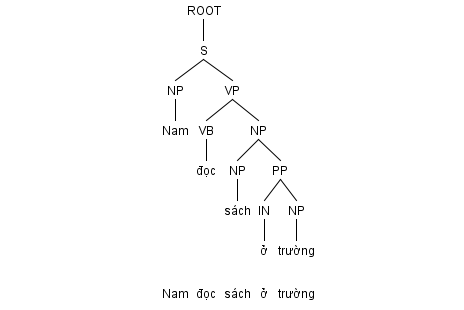
### Hình vẽ cây cú pháp

#### Hình vẽ cây cú pháp 10 câu ngữ liệu huấn luyện

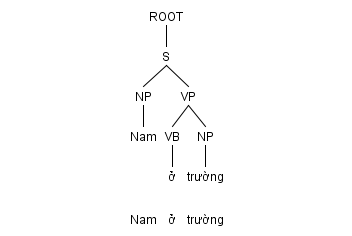
* 1. Nam học bài.



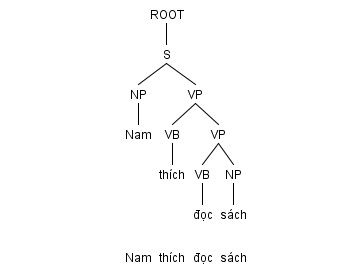
* 1. Nam đọc sách ở trường.



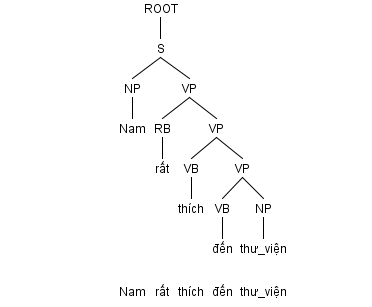
* 1. Nam ở trường.



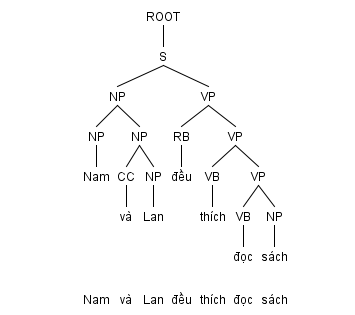
* 1. Nam thích đọc sách.



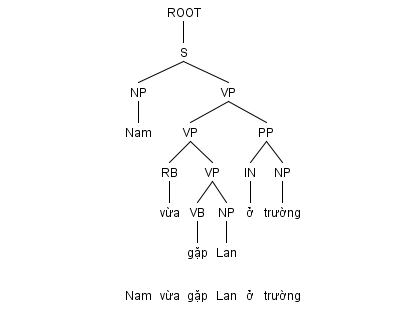
* 1. Nam rất thích đến thư viện.



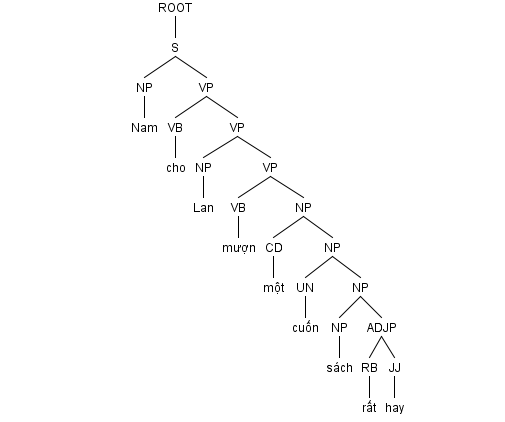
* 1. Nam và Lan đều thích đọc sách.



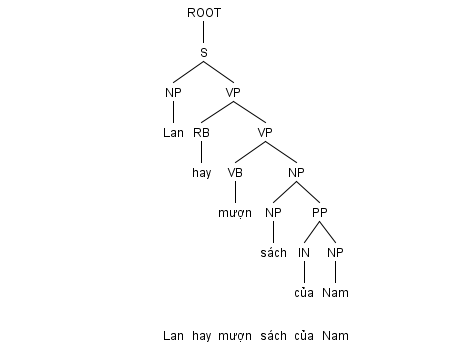
* 1. Nam vừa gặp Lan ở trường.



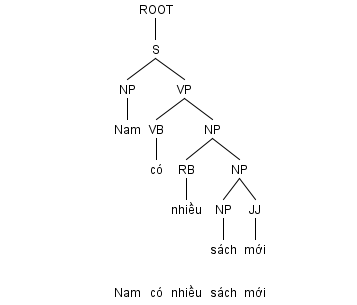
* 1. Nam cho Lan mượn một cuốn sách rất hay.



* 1. Lan hay mượn sách của Nam.

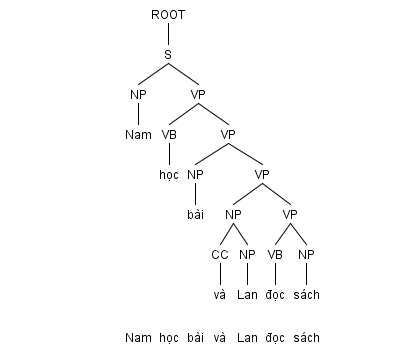


* 1. Nam có nhiều sách mới.

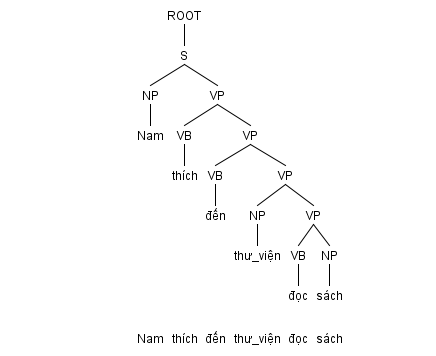


#### Hình vẽ cây cú pháp 10 câu mở rộng

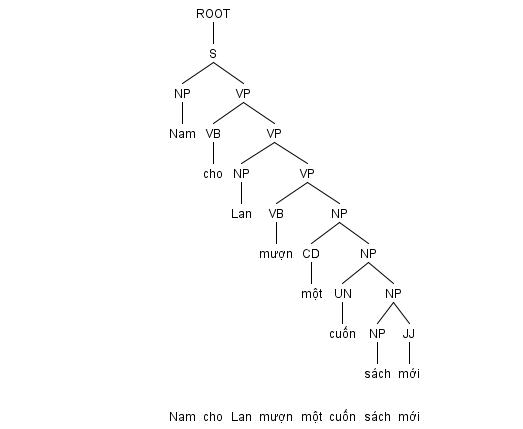
1. Nam học bài và Lan đọc sách.



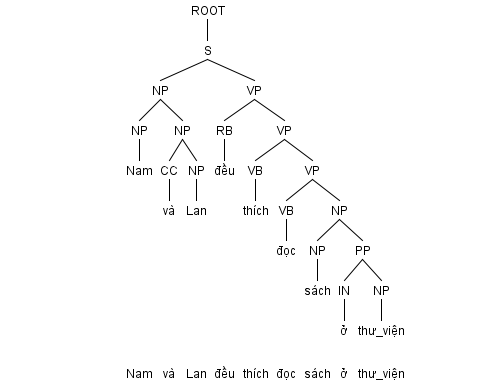
1. Nam thích đến thư viện đọc sách.



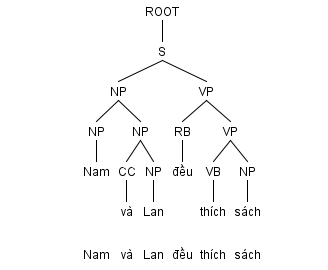
1. Nam cho Lan mượn một cuốn sách mới.



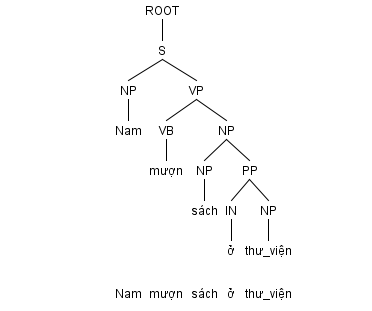
1. Nam và Lan đều thích đọc sách ở thư viện.



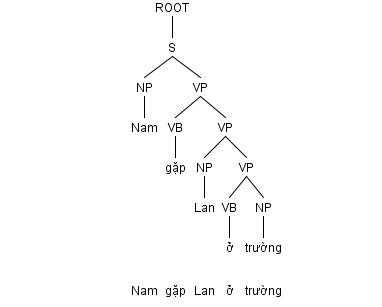
1. Nam và Lan đều thích sách.



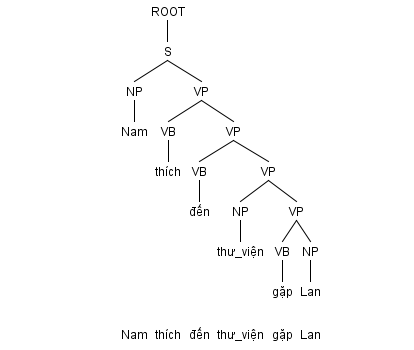
1. Nam mượn sách ở thư viện.



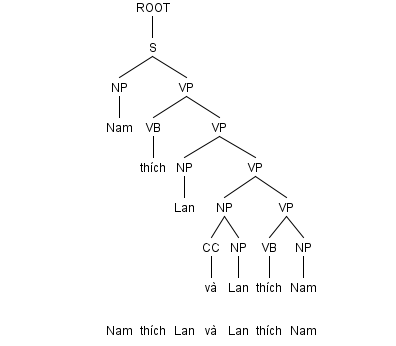
1. Nam gặp Lan ở trường.



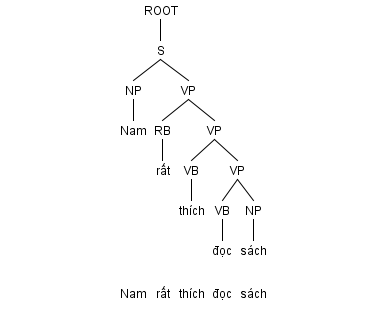
1. Nam thích đến thư viện gặp Lan.



1. Nam thích Lan và Lan thích Nam.

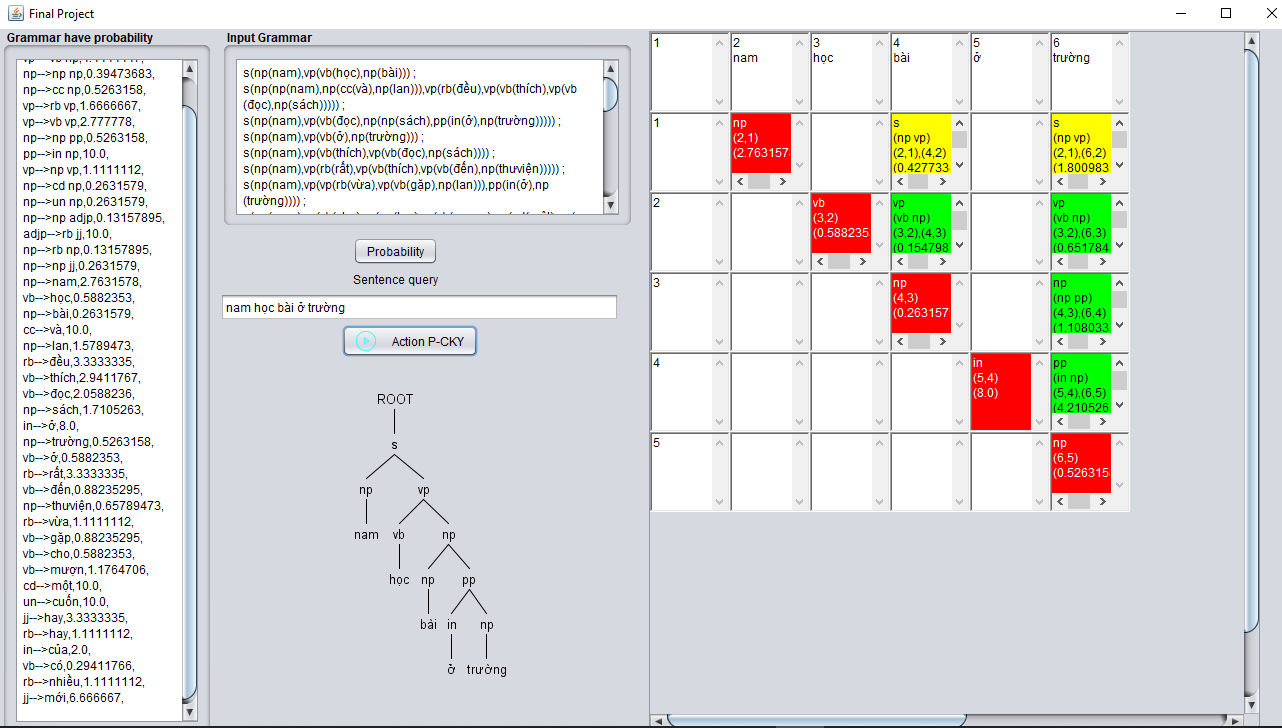


1. Nam rất thích đọc sách.

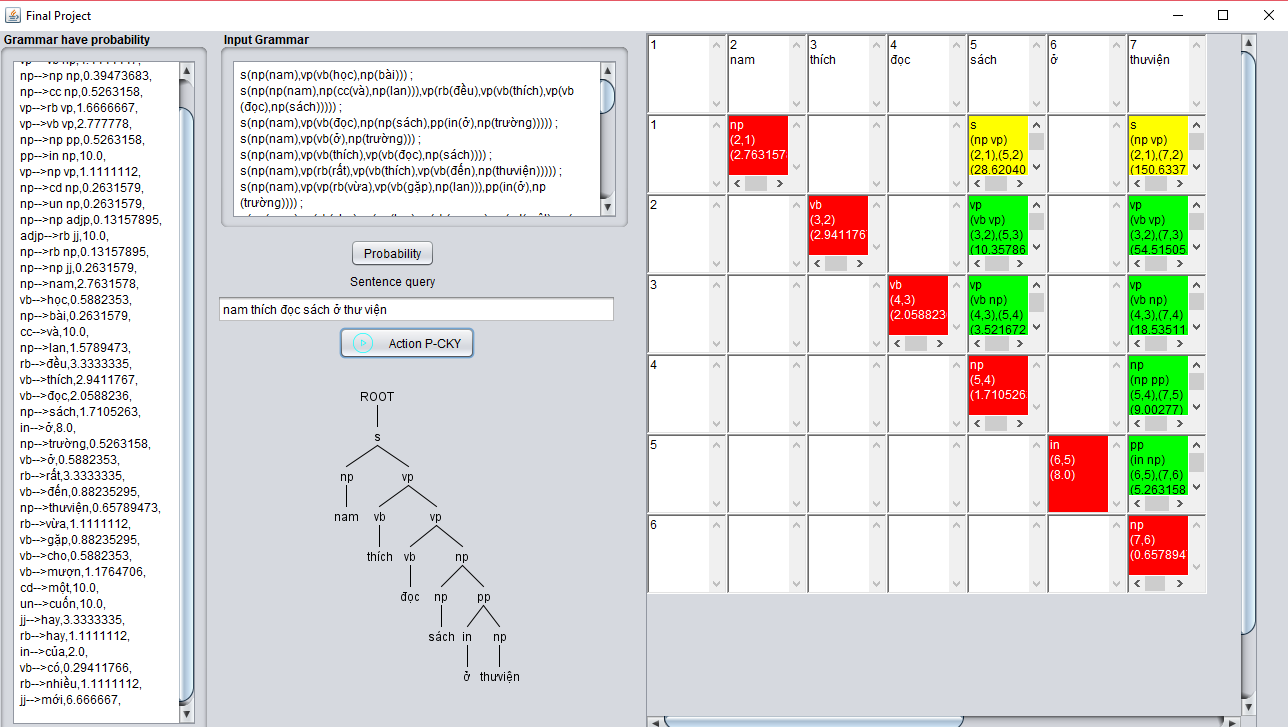


# KẾT QUẢ PHÂN TÍCH CÚ PHÁP BẰNG THUẬT TOÁN PCKY

## Ví dụ 1: Nam học bài ở trường



## Ví dụ 2: Nam rất thích đọc sách ở thư viện.



# tỔNG KẾT

## Kết quả đạt được

* Giao diện dễ nhìn, người dùng dễ sử dụng.
* Chương trình nhận ngữ liệu là cây cú pháp của các câu huấn luyện theo CNF, biểu diễn dưới dạng danh sách.
* Chương trình rút ra các quy tắc cú pháp CNF từ ngữ liệu huấn luyện và tính xác suất cho các quy tắc.
* Chương trình sử dụng thuật toán PCKY để phân tích cú pháp cho các câu thử nghiệm và xuất kết quả.
* Tạo các câu mở rộng và phân tích cú pháp của chúng bằng thuật toán PCKY.
* Vẽ cây cú pháp kết quả.

## Hạn chế

* Chưa vẽ được cây cú pháp kết quả theo 2 nhánh.

Ví dụ: Nam và Lan học bài.

## Hướng phát triển

* Sử dụng thư viện trong bộ công cụ có sẵn để vẽ cây cú pháp kết quả hoàn chỉnh và đẹp hơn.
* Xây dựng phần mềm dễ nhìn, dễ sử dụng và đáp ứng nhu cầu người sử dụng.

# TÀI LIỆU THAM KHẢO

* Bài tập xử lí ngôn ngữ tự nhiên

<https://ohmygodit.wordpress.com/category/bai-tap-xu-ly-ngon-ngu-tu-nhien/>

* WikiPedia : <https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn>

<https://vi.wikipedia.org/wiki/Java_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)>

* Slide hướng dẫn lập trình Java trên Courses của môn Ngôn ngữ lập trình Java.
* **Link Github** : **https://github.com/QuangTrung1104/Op-Ep**