Hướng dẫn sử dụng treebank editor (TBE)

*Nguyễn Phương Thái, Lê Anh Cường*

***Giới thiệu***

TBE là công cụ trợ giúp người làm dữ liệu gán nhãn cho câu ở nhiều mức độ khác nhau (tách từ, gán nhãn từ loại, xây dựng cây cú pháp). Các sửa đổi trên dữ liệu sẽ được lưu lại dạng file log (để thống kê, theo dõi tiến độ). Hiện tại TBE chưa được tích hợp các công cụ phân tích tự động[[1]](#footnote-1).

Contents

[I. Các chế độ làm việc 1](#_Toc235513996)

[II. Soạn thảo cây 2](#_Toc235513997)

[III. Gán nhãn từ loại 3](#_Toc235513998)

[IV. Tách từ 4](#_Toc235513999)

[V. Xem file log 6](#_Toc235514000)

[V.1 Xem log cây cú pháp 6](#_Toc235514001)

[V.2 Xem log gán nhãn từ loại 6](#_Toc235514002)

[VI. Tìm kiếm và thống kê 7](#_Toc235514003)

[VII. Chế độ làm việc kiểm tra đồng thuận 8](#_Toc235514004)

[VIII. Tab Câu và tab Tách câu 8](#_Toc235514005)

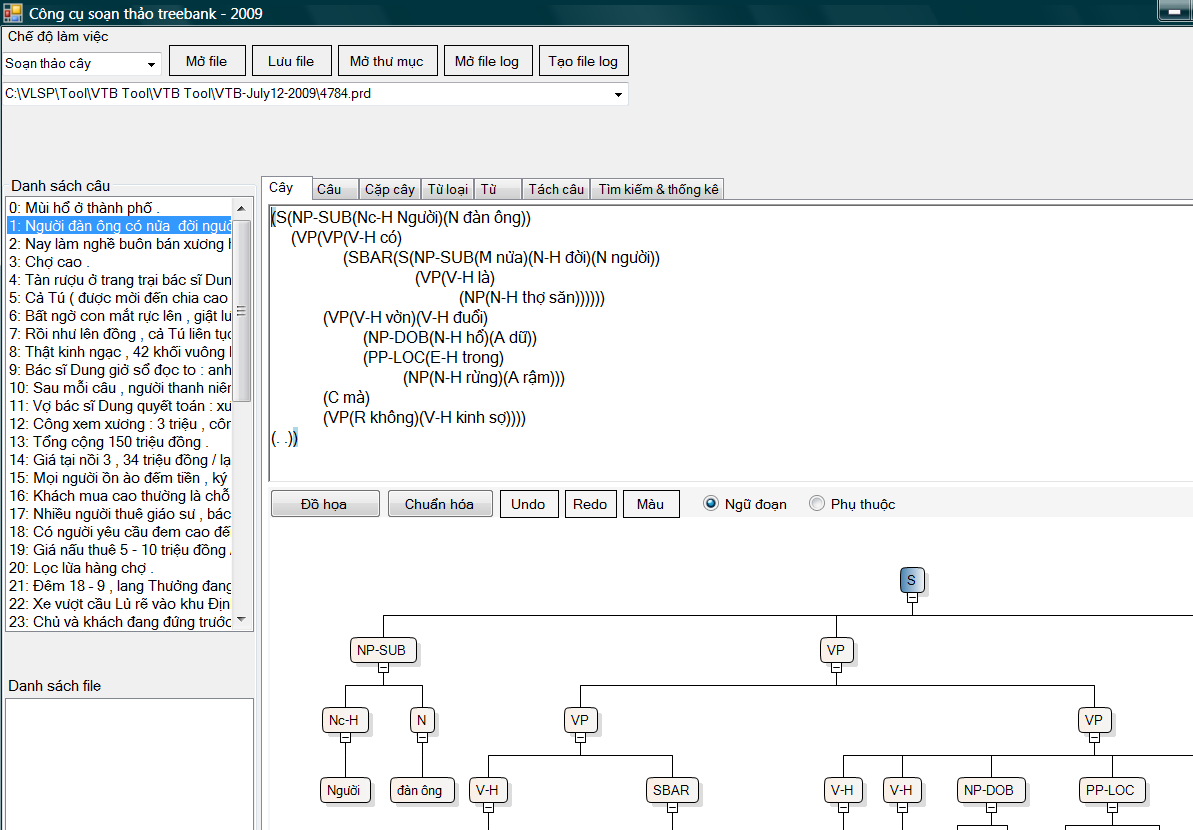
# Các chế độ làm việc

Có bảy chế độ làm việc:

* Soạn thảo cây (\*)
* Gán nhãn từ loại (\*)
* Tách từ (\*)
* Xem log cây
* Xem log từ loại
* Xem log tách từ
* Kiểm tra đồng thuận

Việc chọn chế độ làm việc được yêu cầu thực hiện ngay khi chạy TBE. Các chế độ có dấu \* (chế độ soạn thảo) sẽ ghi ra file log các sửa đổi trên dữ liệu của người làm dữ liệu. Do đó người dùng sẽ phải tạo file log mới hoặc mở file log đã có trước khi mở file dữ liệu.

# Soạn thảo cây



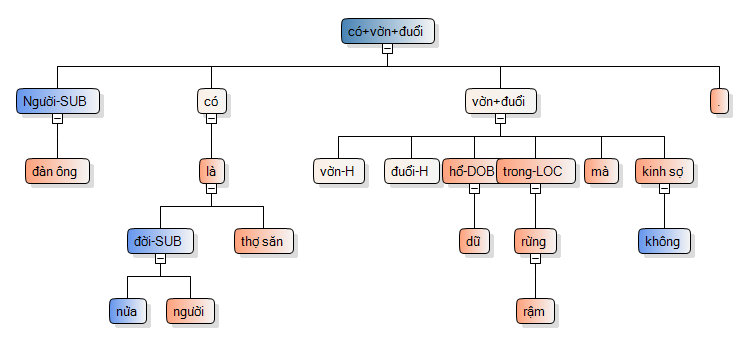
Hình . Chế độ soạn thảo cây cú pháp với file dữ liệu mở

Hình 1 minh họa TBE ở chế độ soạn thảo cây cú pháp. Ở góc trái trên là combo box chứa danh sách chế độ làm việc và hiện tại đang là “Soạn thảo cây”. Ngay bên dưới là combo box khác chứa tên file đang mở “C:\VLSP\Tool\VTB Tool\VTB Tool\VTB-July12-2009\4784.prd”. Dưới nữa là ba cửa sổ quan trọng khác:

1. Cửa sổ danh sách câu (bên trái): liệt kê các câu trong file dữ liệu đang mở
2. Cửa sổ hiển thị cây ở chế độ văn bản (phải-trên): cho phép người làm dữ liệu soạn thảo cây. Các cặp dấu ngoặc tương ứng với nhau sẽ được highlight.
3. Cửa sổ hiển thị cây ở chế độ đồ họa (phải-dưới)

Giữa hai cửa sổ hiển thị cây là một số nút:

* Nút “Đồ họa”: tạo cây đồ họa ở cửa sổ (3) tương ứng với câu text đang có trong cửa sổ (2)
* Nút “Chuẩn hóa”: Chuẩn hóa qui cách (nút con phải lùi vào so với nút cha, các nút con phải thẳng hàng nhau, v.v.) cho cây dạng text ở cửa sổ (2)
* Nút “Undo”: hủy bỏ các hành động soạn thảo gần đây nhất
* Nút “Redo”: khôi phục các hành động soạn thảo gần đây nhất
* Nút “Màu”: Chọn màu đen cho văn bản ở cửa sổ (2)
* Nút “Ngữ đoạn”: Cây đồ họa sẽ ở dạng ngữ đoạn
* Nút “Phụ thuộc”: Cây đồ họa sẽ ở dạng phụ thuộc (xem ví dụ ở Hình 2). Cây phụ thuộc là cây mà nhãn của nút là từ (có thể là một hoặc nhiều) chứ không phải phân loại cú pháp, nhãn ở nút con là từ phụ thuộc vào từ ở nút cha, màu xanh để chỉ từ bên trái, màu hồng chỉ từ bên phải (của từ trung tâm ở nút cha).

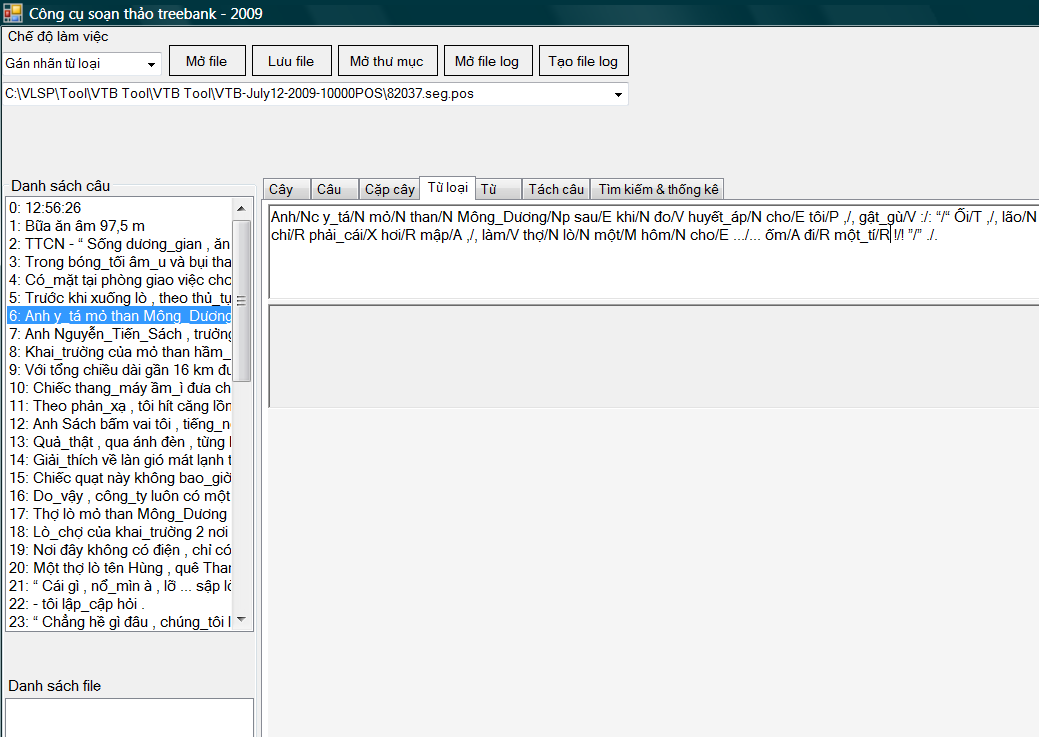


Hình . Cây phụ thuộc của câu trong Hình 1

**Qui trình làm việc:**

* Chọn chế độ “Soạn thảo cây”
* Mở hoặc tạo file log (nút “Mở file log” hoặc nút “Tạo file log”)
* Mở file dữ liệu có tên dạng \*.prd (nút “Mở file”)
* Chọn tab “Cây” (mặc định)
* Sửa cây cú pháp ở cửa sổ (2)
* Ấn "Lưu file" sau khi sửa mỗi cây

# Gán nhãn từ loại

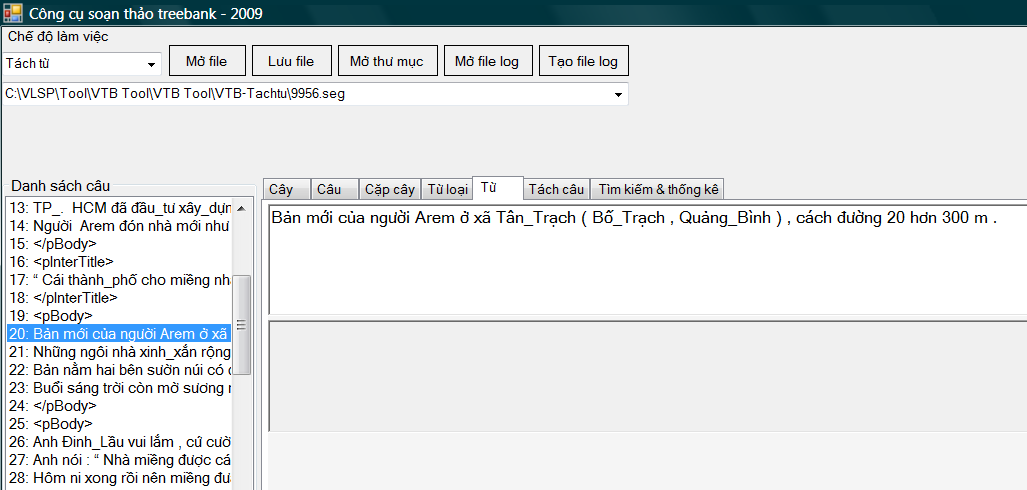


Hình . Chế độ gán nhãn từ loại với file dữ liệu đang mở

**Qui trình làm việc:**

* Chọn chế độ “Gán nhãn từ loại”
* Mở hoặc tạo file log (nút “Mở file log” hoặc nút “Tạo file log”)
* Mở file dữ liệu có tên dạng \*.pos (nút “Mở file”)
* Chọn tab “Từ loại”
* Gán nhãn từ loại cho câu ở cửa sổ thuộc tab “Từ loại”
* Ấn "Lưu file" sau khi sửa mỗi câu

# Tách từ



Hình . Chế độ tách từ với file dữ liệu đang mở

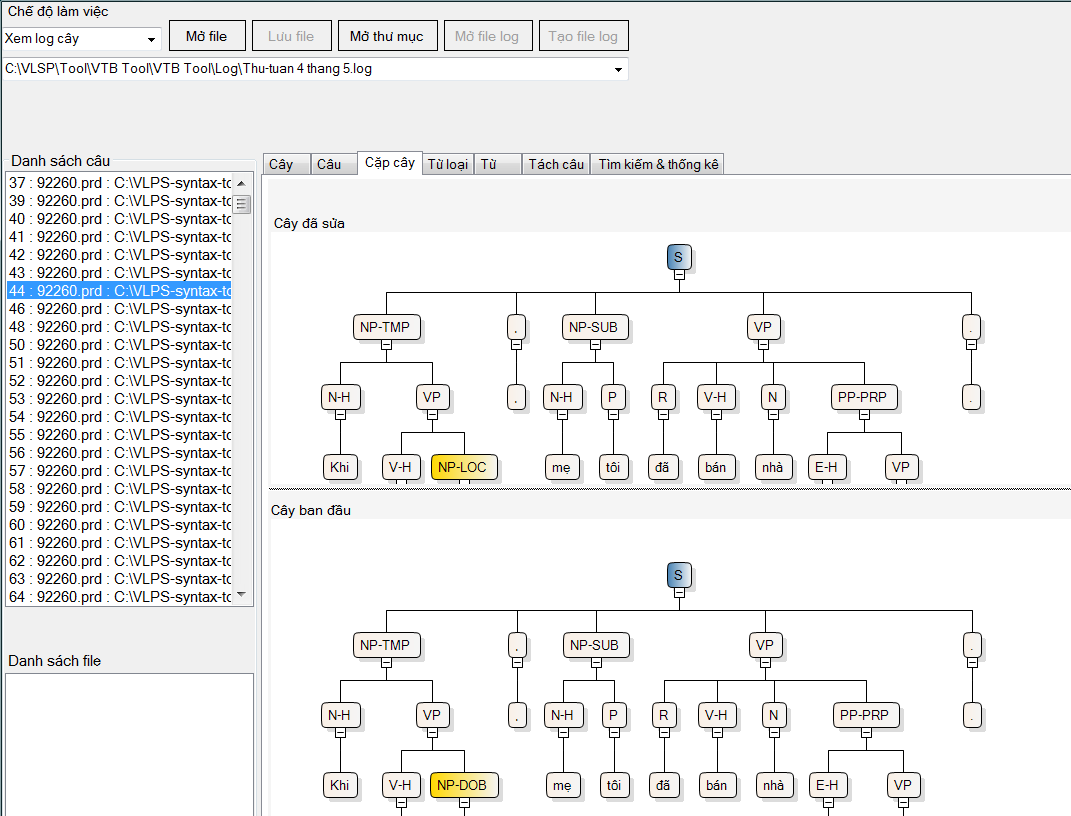
**Qui trình làm việc:**

* Chọn chế độ làm việc là “Tách từ”
* Mở hoặc tạo file log (nút “Mở file log” hoặc nút “Tạo file log”)
* Mở file dữ liệu có tên dạng \*.seg (nút “Mở file”)
* Chọn tab “Từ”
* Gán nhãn từ loại cho câu ở cửa sổ thuộc tab “Từ”
* Ấn "Lưu file" sau khi sửa mỗi câu

Chú ý : Trong danh sách câu có cả các nhãn XML (ví dụ </pBody>, <pInterTitle>) thì bỏ qua, không sửa gì các nhãn đó.

# Xem file log

## V.1 Xem log cây cú pháp



Hình . Xem log cây

**Qui trình làm việc:**

* Chọn chế độ làm việc “Xem log cây”
* Mở file log có tên dạng \*.log (nút “Mở file”)
* Chọn tab “Cặp cây”
* Xem danh sách câu đã được sửa (nhắp đúp chuột vào từng câu trong danh sách câu): Ở tab “Cặp cây” có hai cửa sổ là “Cây đã sửa” và “Cây ban đầu”. Các nút được sửa sẽ có màu vàng, nút mới hoàn toàn thì màu xanh (ở Hình 5 có một nút vàng là NP-LOC trong cửa sổ “Cây đã sửa”).

## V.2 Xem log gán nhãn từ loại

**Qui trình làm việc:**

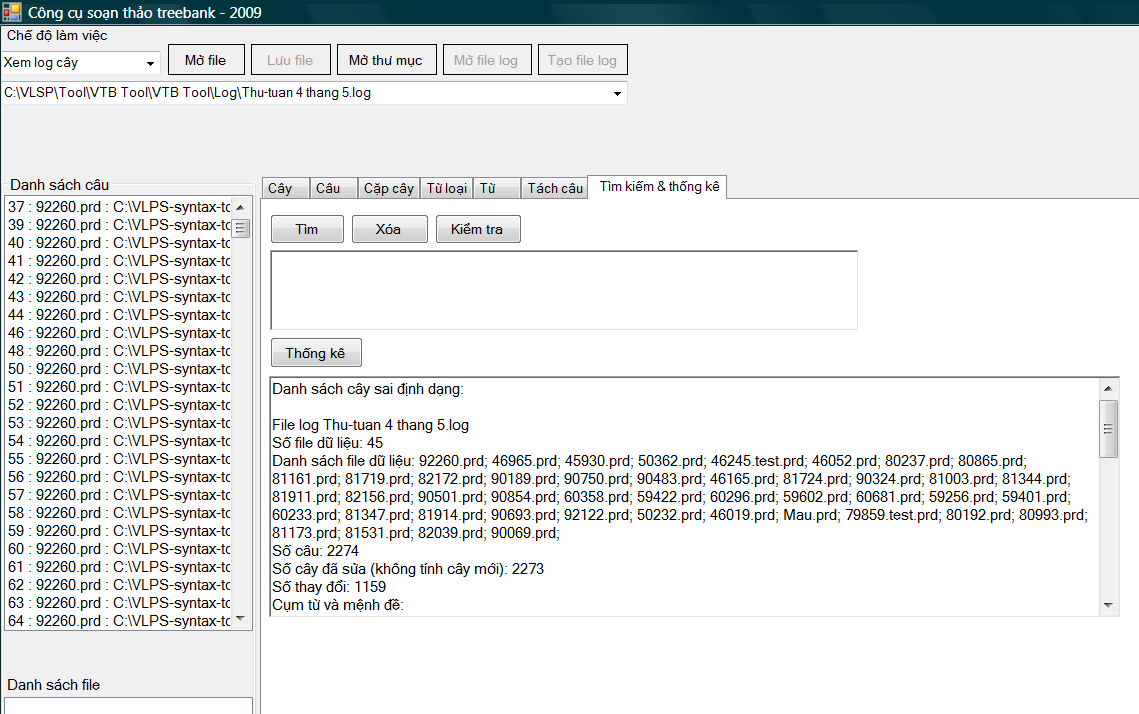
* Chọn chế độ làm việc “Xem log từ loại”
* Mở file log có tên dạng \*.log (nút “Mở file”)
* Chọn tab “Từ loại”
* Xem danh sách câu đã được sửa (nhắp đúp chuột vào từng câu trong danh sách câu): Ở tab “Từ loại” có hai cửa sổ, cái trên hiển thị câu đã sửa (từ màu đỏ là đã được sửa), cái dưới hiển thị câu gốc.

# Tìm kiếm và thống kê

Đây là tính năng đi kèm với các chế độ soạn thảo (“Soạn thảo cây”, “Gán nhãn từ loại”, “Tách từ”). Ta có thể tìm:

* Theo từ: ở cả 3 chế độ soạn thảo
* Theo nhãn thành phần cú pháp, nhãn nút con của nó, và từ trung tâm: chỉ dành cho chế độ “Soạn thảo cây”. Ví dụ:
  + Nếu khóa là “t=PP-LOC” thì các cây chứa nhãn PP-LOC sẽ được tìm ra
  + Nếu khóa là “t=VP;ct=SBAR” thì các cây có cụm VP mà con của nó là SBAR sẽ được tìm ra
  + Nêu khóa là “t=VP;hw=lấy” thì các cây có cụm VP mà trung tâm là từ “lấy” sẽ được tìm ra

Các câu được tìm thấy sẽ được đánh dấu “\*” trong danh sách câu.



Hình . Thống kê file log cây

Tính năng thống kê khi xem log cây (Hình 6) sẽ đưa ra cho chúng ta:

* Danh sách file đã sửa
* Số nút mới, bị sửa, bị xóa (nút có thể là từ, từ loại, hay cụm từ và câu)

# Chế độ làm việc kiểm tra đồng thuận

Ở chế độ làm việc này, qui trình làm việc như sau:

* Chọn chế độ làm việc “Kiểm tra đồng thuận”
* Mở file của người thứ nhất
* Mở file của người thứ hai
* Chọn tab “Cặp cây”
* Xem các cặp cây giống chế độ “Xem log cây cú pháp”

# Tab Câu và tab Tách câu

Tab Câu hiển thị toàn bộ các câu trong văn bản đang xét và highlight câu hiện thời. Người làm dữ liệu có thể dùng tab này trong các chế độ soạn thảo.

Tab Tách câu dùng để tách lại câu và hoạt động với chế độ Gán nhãn từ loại và Tách từ. Người làm dữ liệu sẽ trực tiếp soạn thảo trên file nguồn được mở trong cửa sổ thuộc tab Tách câu. Họ cần chú ý giữ đúng khuôn dạng file. Tách xong cần ấn nút Lưu file ngay.

1. Các công cụ này sẽ được dùng để tiền xử lý dữ liệu [↑](#footnote-ref-1)