CLC PARALLEL CORPUS TOOL V2.0

08/2016

COMPUTATIONAL LINGUISTICS CENTER
227 Nguyễn Văn Cừ, Q5, TPHCM

Mục lục

GIỚI THIỆU

CLC Parallel Corpus Tool là công cụ dùng để khai thác ngữ liệu song ngữ phục vụ cho nhiều mục đích khác nhau (giảng dạy, nghiên cứu, …), đối chiếu từ vựng trong tiếng Anh sang tiếng Việt và ngược lại, cũng như hỗ trợ thống kê từ vựng trong ngữ liệu dựa trên các tiêu chí nhất định và thông tin nhãn kèm theo.

Chương trình ngoài hỗ trợ tiếng Anh còn hỗ trợ tiếng Hoa và tiếng Hàn.

YÊU CẦU HỆ THỐNG

Chương trình yêu cầu phải có .NET Framework 4.0 trở lên (tương đương từ hệ điều hành Windows 8 trở về sau).

HƯỚNG DẪN SỬ DỤNG

1. Nạp ngữ liệu

Ta nạp ngữ liệu cho chương trình bằng một trong các cách sau:

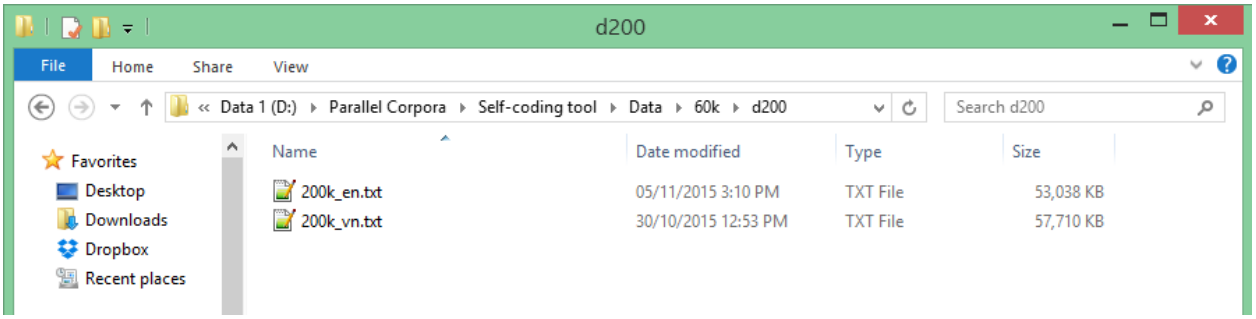
- Ấn vào nút thư mục trên thanh công cụ.
- Vào menu File / Open hoặc dùng tổ hợp phím Ctrl + O.

Hình 1: Các nút chức năng của chương trình.

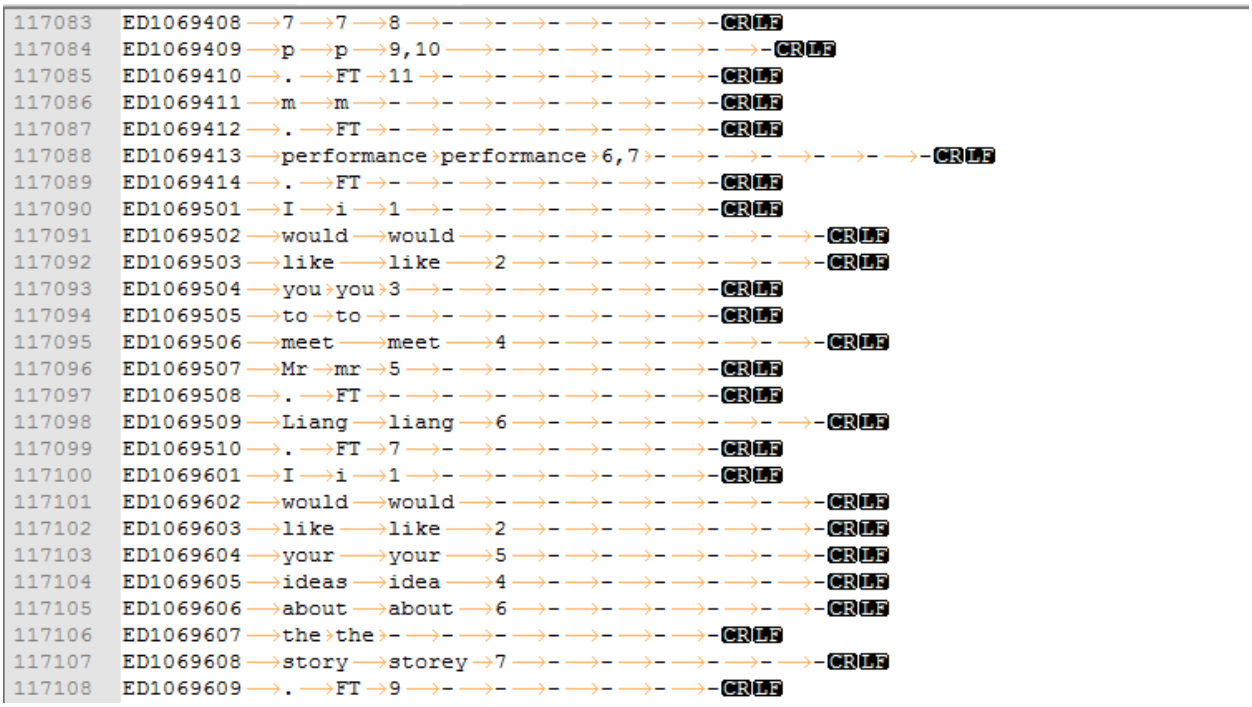
Về ngữ liệu đầu vào:

- Tên file phải có định dạng: <tên bất kì>_<ngôn ngữ>.txt. Ví dụ: 60k_en.txt, 60k_vn.txt, …
- Trường hợp load song ngữ, ta phải có 2 file, mỗi file là 1 ngôn ngữ, được đặt cùng tên phần <tên bất kì> và đặt cùng 1 thư mục.
- Trường hợp load đơn ngữ, ta chỉ cần 1 file nhưng tên file bắt buộc phải có định dạng như trên để chương trình vẫn có thể hiểu được.
- File data bắt buộc có định dạng 10 cột. Mỗi cột cách nhau bằng 1 dấu tab. Ý nghĩa các cột theo thứ tự từ trái sang phải như sau:
  - ID: mã số của từng từ, cho biết từ đó là từ thứ mấy nằm trong câu và cho biết câu đang chứa từ đó là câu thứ mấy trong data.
  - Word: từ

- Lemma: tương tự như cột Word nhưng các từ là danh từ, động từ sẽ được biểu diễn bằng hình thái gốc của nó trong từ điển.
- Links: thông tin gióng hàng qua ngôn ngữ còn lại.
- Morpho: tất cả các từ (ngoại trừ số, tên riêng và các kí hiệu đặc biệt như $, ^, &, …) sẽ được biểu diễn bằng hình thái gốc của nó trong từ điển.
- POS: nhãn từ loại của từ.
- Phrase: cụm từ mà từ này nằm trong (thường là nằm trong 1 chunk)
- Grm: thông tin về ngữ pháp của từ trong câu
- NER: nhãn thực thể có tên
- Semantic: nhãn ngữ nghĩa (hiện tại đang sử dụng nhãn của LLOCE)



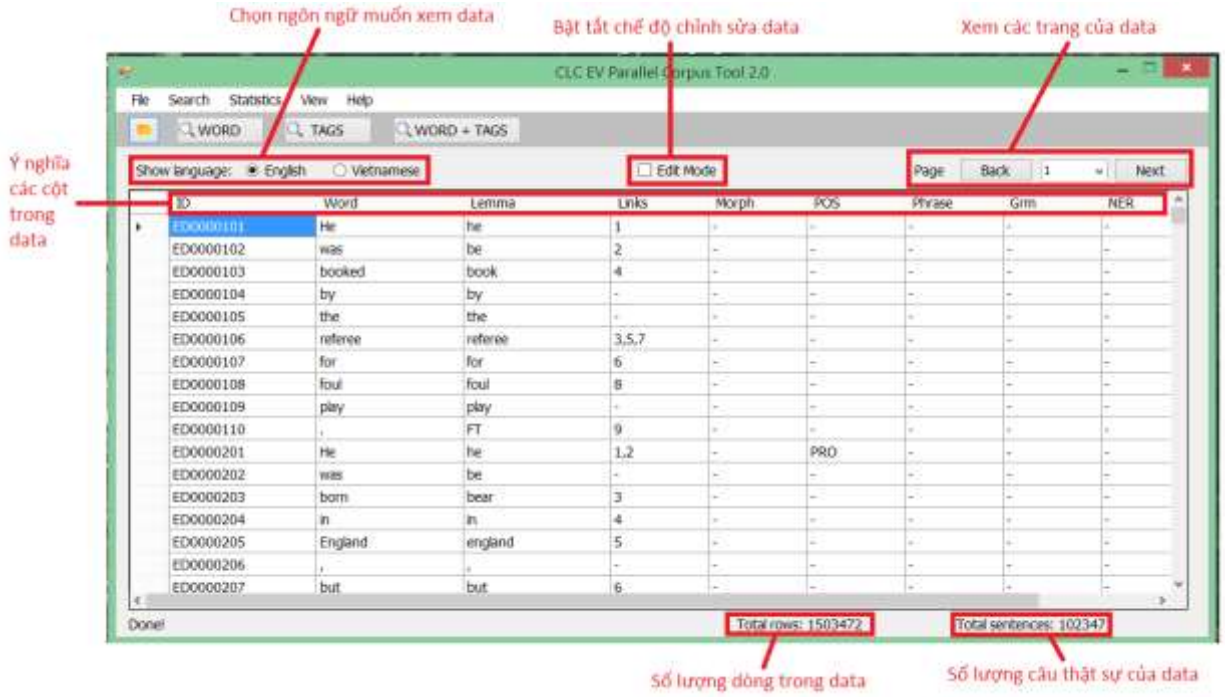
Hình 2: Ví dụ về tên tập tin.



Hình 3: Định dạng của data đầu vào.

Ta chọn trong phần Show Language ngôn ngữ muốn xem để chương trình hiển thị.

Do ngữ liệu lớn nên sẽ được chia ra hiển thị theo trang, mỗi trang hiển thị tối đa 500 dòng trong ngữ liệu.

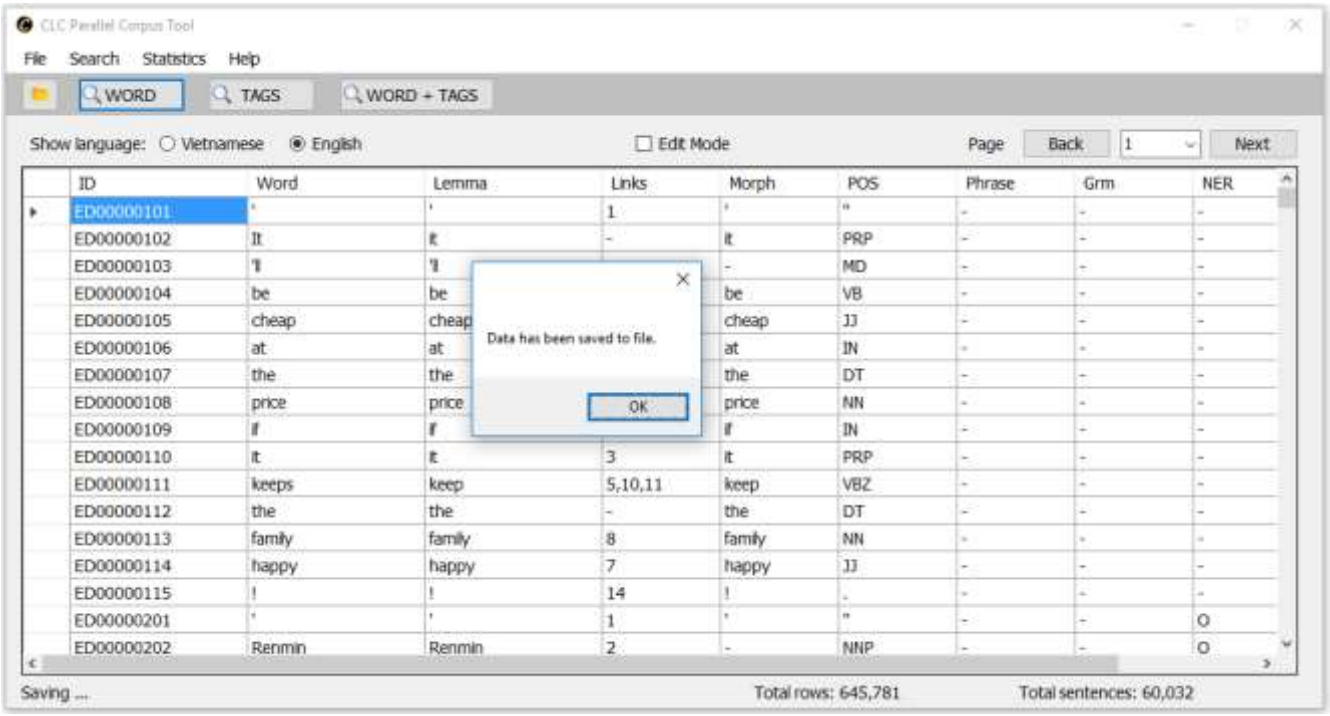


Hình 4: Chi tiết các phần hiển thị của chương trình.

2. Chỉnh sửa ngữ liệu

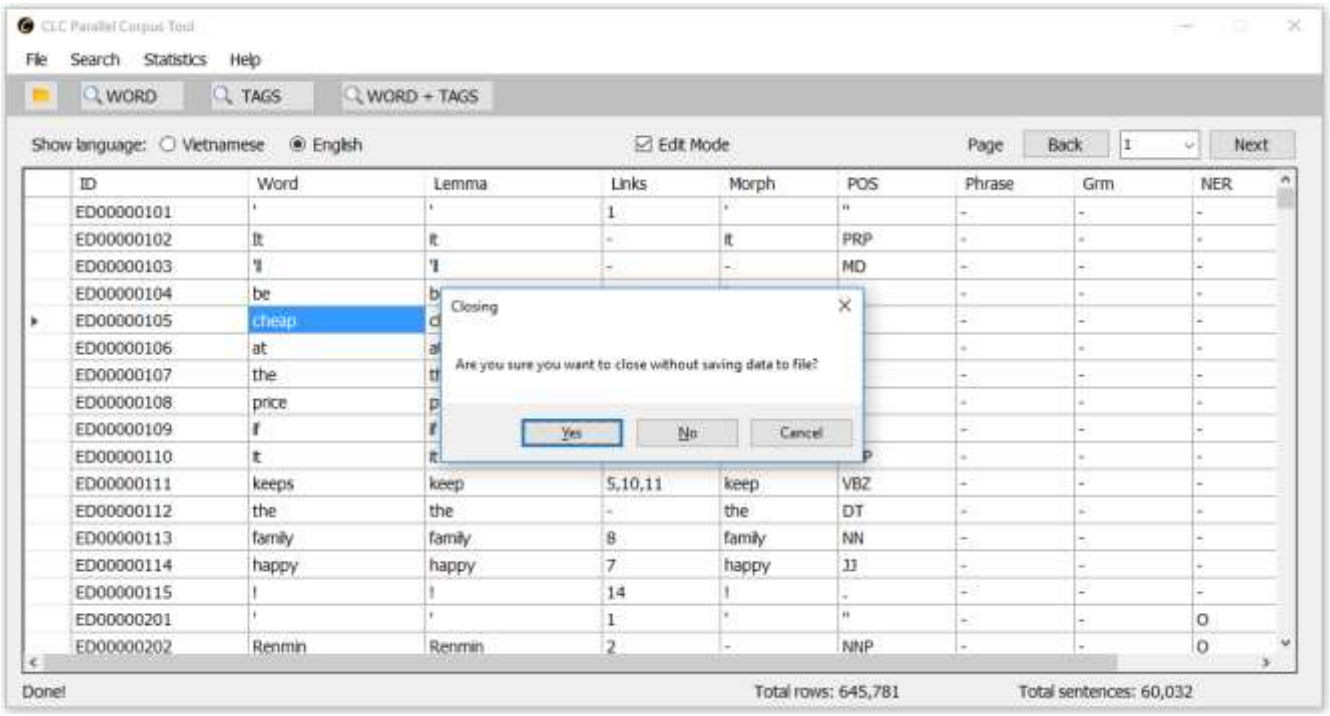
Để chỉnh sửa ngữ liệu, ta chọn vào chức năng Edit Mode. Nếu không muốn chỉnh sửa nữa, ta chỉ cần bỏ chọn chức năng này.

Lưu ý lúc này ngữ liệu được chỉnh sửa chỉ ảnh hưởng trong chương trình, chưa ảnh hưởng đến file ngữ liệu gốc, nếu muốn lưu lại những gì đã chỉnh sửa, ta cần bấm vào File/Save hoặc tổ hợp phím Ctrl + S để lưu lại. Khi lưu thành công, sẽ có bảng thông báo hiện lên.



Hình 5: Thông báo khi lưu ngữ liệu thành công

Trường hợp ta có chỉnh sửa ngữ liệu nhưng không muốn lưu mà muốn thoát khỏi chương trình, chương trình sẽ hiện thông báo để xác minh với ta thêm 1 lần nữa. Khi hiện thông báo này, nếu nhấn Yes thì chương trình sẽ tự động tắt mà không lưu gì cả, nhấn No thì chương trình sẽ tự động lưu ngữ liệu lên file gốc và tự động tắt, còn nhấn Cancel thì vẫn tiếp tục ở lại chương trình.



Hình 6: Bảng thông báo xác nhận lần nữa
khi tắt chương trình nếu trước đó chưa lưu ngữ liệu.

Lưu ý: Chỉnh sửa ngữ liệu được xác định là khi ta đã thay đổi giá trị của 1 ô bất kì, cho dù đó là giá trị nào đi nữa và cho dù sau khi ta bỏ chọn chức năng Edit Mode thì chương trình vẫn ngầm hiểu là ta đã thay đổi giá trị, nên ta nhớ lưu lại hoặc khi tắt chương trình sẽ hỏi để xác nhận lại lần nữa việc sao lưu các ngữ liệu đã thay đổi.

3. Tìm kiếm

Có 3 kiểu tìm kiếm: word, tags, word + tags.

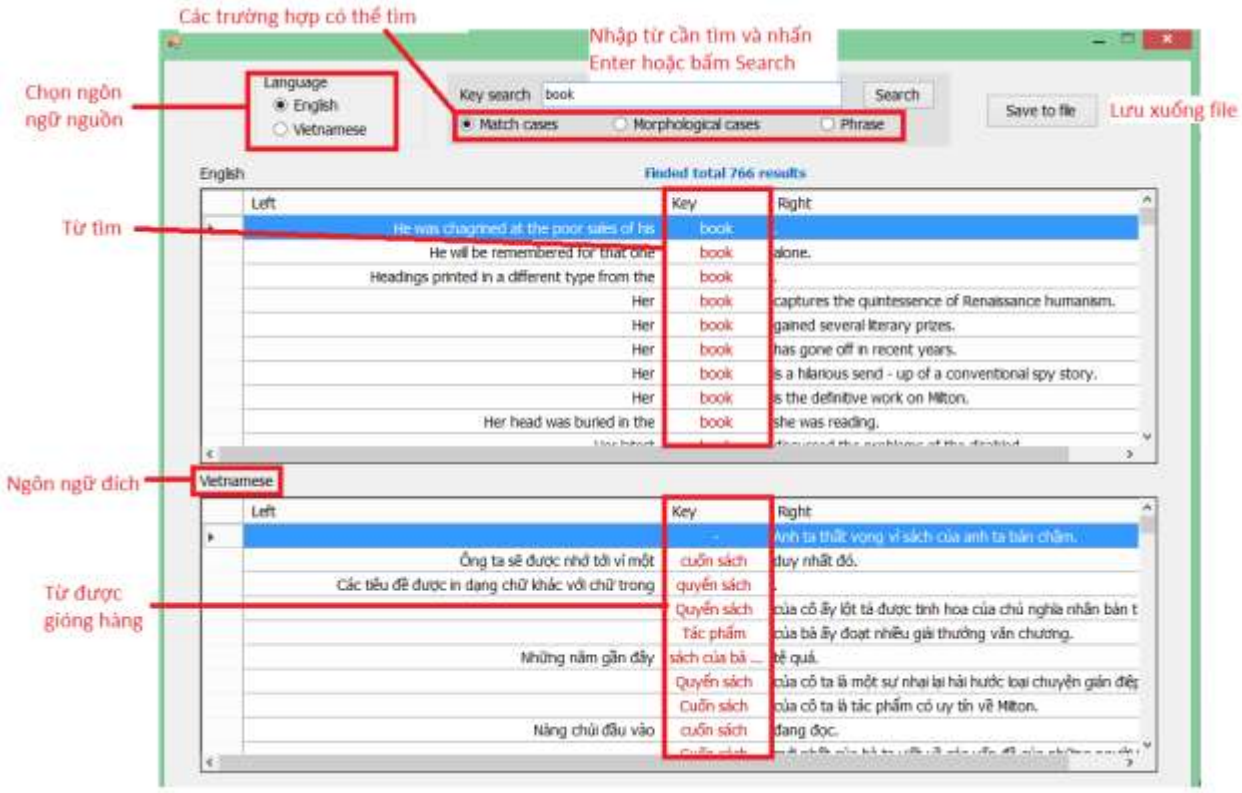
Bấm vào nút có hiển thị tên tương ứng để vào chức năng tìm kiếm.

Cách tìm:

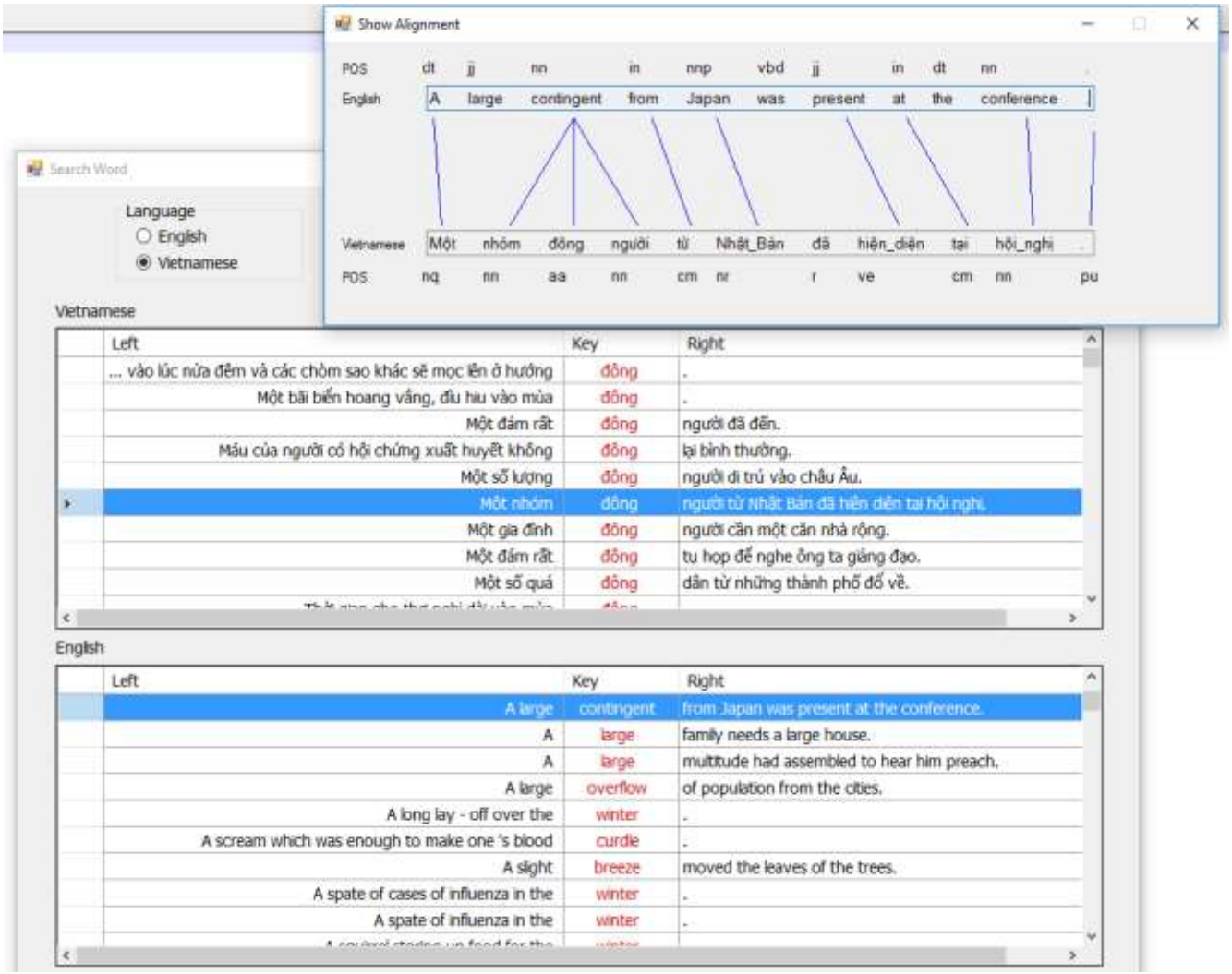
- Đối với Word:
  - Trong cửa sổ tìm kiếm, ta sẽ thấy khung cho chọn ngôn ngữ, đây là ngôn ngữ mà ta sẽ nhập vào trong khung Key Search để tìm. Ta có thể chọn ngôn ngữ mình muốn gõ và gõ từ cần tìm, sau đó nhấn Enter (hoặc bấm nút Search) để chương trình tiến hành tìm kiếm.
  - Có 3 chức năng mở rộng cho tìm kiếm Word, đó là Match cases, Morphological cases hay Phrase.

- ✓ Match cases: chương trình sẽ tự tìm đúng từ mà ta gõ trong khung key search, đồng thời tìm từ được gióng hàng với từ này trong ngôn ngữ còn lại.
- ✓ Morphological cases: chương trình tìm các từ có cùng hình thái gốc là từ trong khung tìm kiếm mà ta nhập vào, lưu ý không phải là tìm hình thái gốc của từ nhập trong khung tìm kiếm.
- ✓ Phrase: cho phép ta nhập vào 1 chuỗi từ, chương trình sẽ tìm kiếm chuỗi từ và trả ra kết quả gióng hàng ở ngôn ngữ đích với từ đầu tiên trong phrase của ngôn ngữ nguồn.

- Trong cửa sổ hiện kết quả, ngôn ngữ nguồn (ngôn ngữ ta chọn nhập trong khung tìm kiếm) sẽ hiển thị ở trên, ngôn ngữ đích sẽ hiển thị ở dưới.
- Khung kết quả sẽ có dạng 3 cột. Phần Left là phần đứng trước từ được gióng hàng. Phần Key là từ tìm kiếm (trong ngôn ngữ nguồn) và từ được gióng hàng (trong ngôn ngữ đích). Phần Right là phần còn lại sau từ cần tìm (từ được gióng hàng).
- Còn có chức năng save file, save lại kết quả mà ta đã tìm kiếm ra file ***.txt***. Cụ thể file sẽ được lưu cùng thư mục chứa chương trình và để cho dễ nhớ, tên file sẽ được đặt theo dạng ***<từ cần tìm>_<dạng tìm kiếm>.txt***.
- Ngoài ra khi nhấp đúp vào 1 câu bất kì, ta có thể xem được chi tiết phần gióng hàng của câu đó.



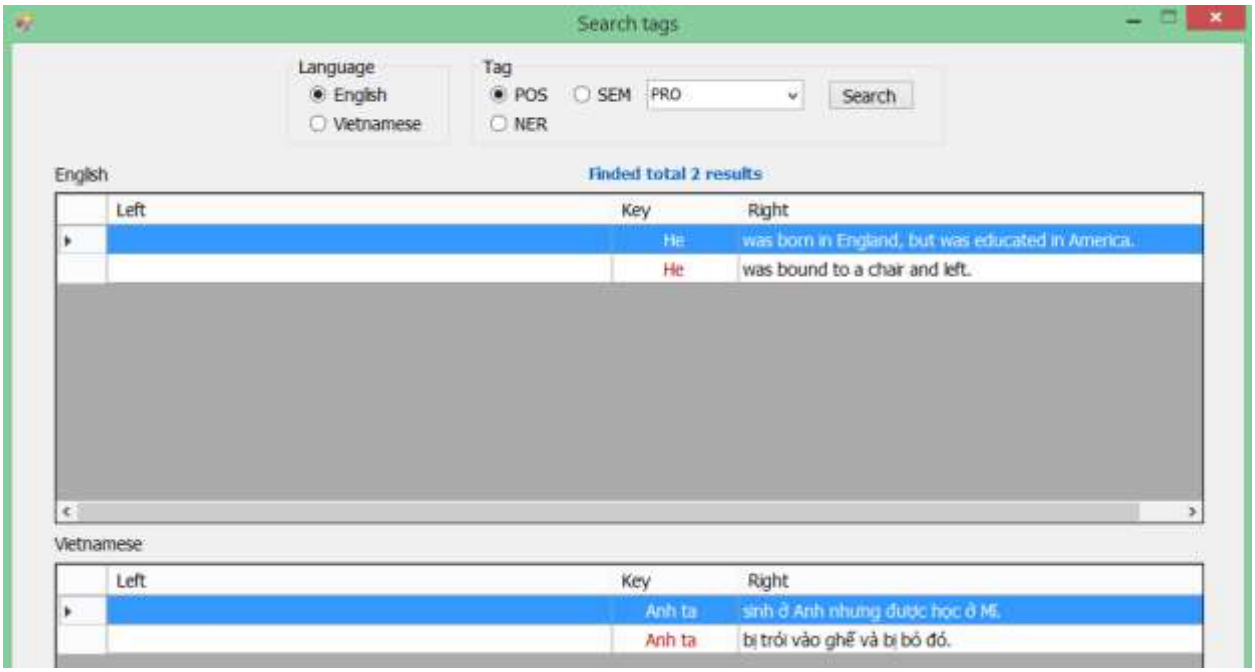
Hình 7: Tổng quan cửa sổ tìm kiếm.



Hình 8: Hiển thị chi tiết phần gióng hàng.

Lưu ý: độ chính xác gióng hàng phụ thuộc vào ngữ liệu đầu vào, đồng nghĩa với khả năng gióng hàng của công cụ gióng hàng (GIZA++). Tương tự với việc gán nhãn từ loại, nhãn thực thể và các loại nhãn khác. Do việc gán nhãn là tự động, không thể chính xác hoàn toàn 100% nên kết quả hiển thị chắc chắn sẽ có vài sai sót.

- Đối với Tag:
  - Ta chọn nhãn mà muốn tìm, và chương trình sẽ trả về từ đầu tiên có nhãn đó trong câu.



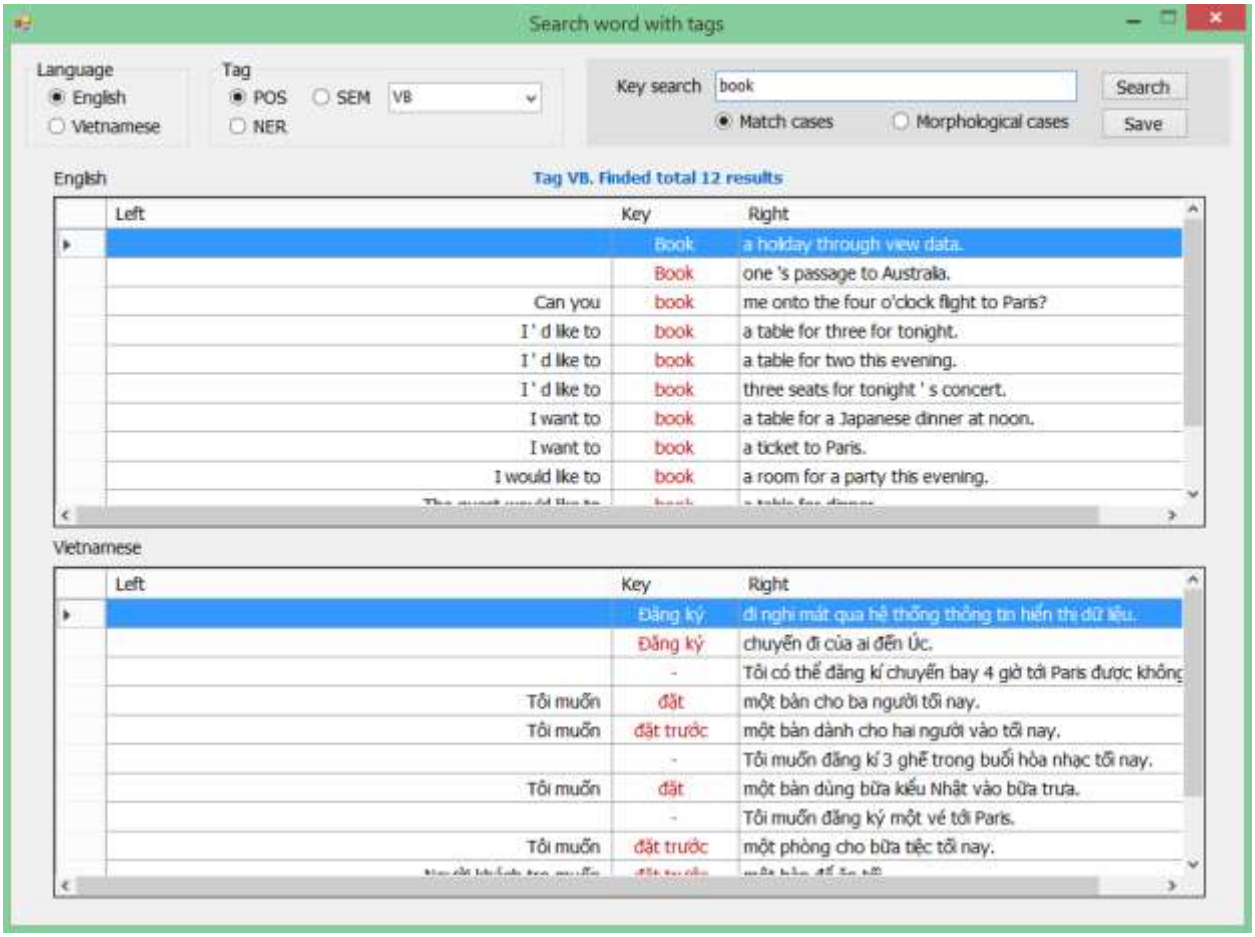
Hình 9: Ví dụ về tìm kiếm nhãn.

- Đối với Word + Tags:
  - Là sự kết hợp của cả 2 loại tính năng kể trên: vừa tìm kiếm từ theo 1 dạng nhất định nào đó (Match cases hay Morphological cases) kết hợp với các nhãn bổ sung thêm thông tin cho từ cần tìm.
  - Lưu ý ở đây do có nhãn bổ sung thêm thông tin nên chương trình chỉ tìm kiếm ở dạng từ, chưa được thiết kế để tìm kiếm ở dạng Phrase.
  - File khi lưu xuống sẽ có định dạng:

<từ cần tìm>_<dạng tìm kiếm>_<loại nhãn>_<tên nhãn>.txt.

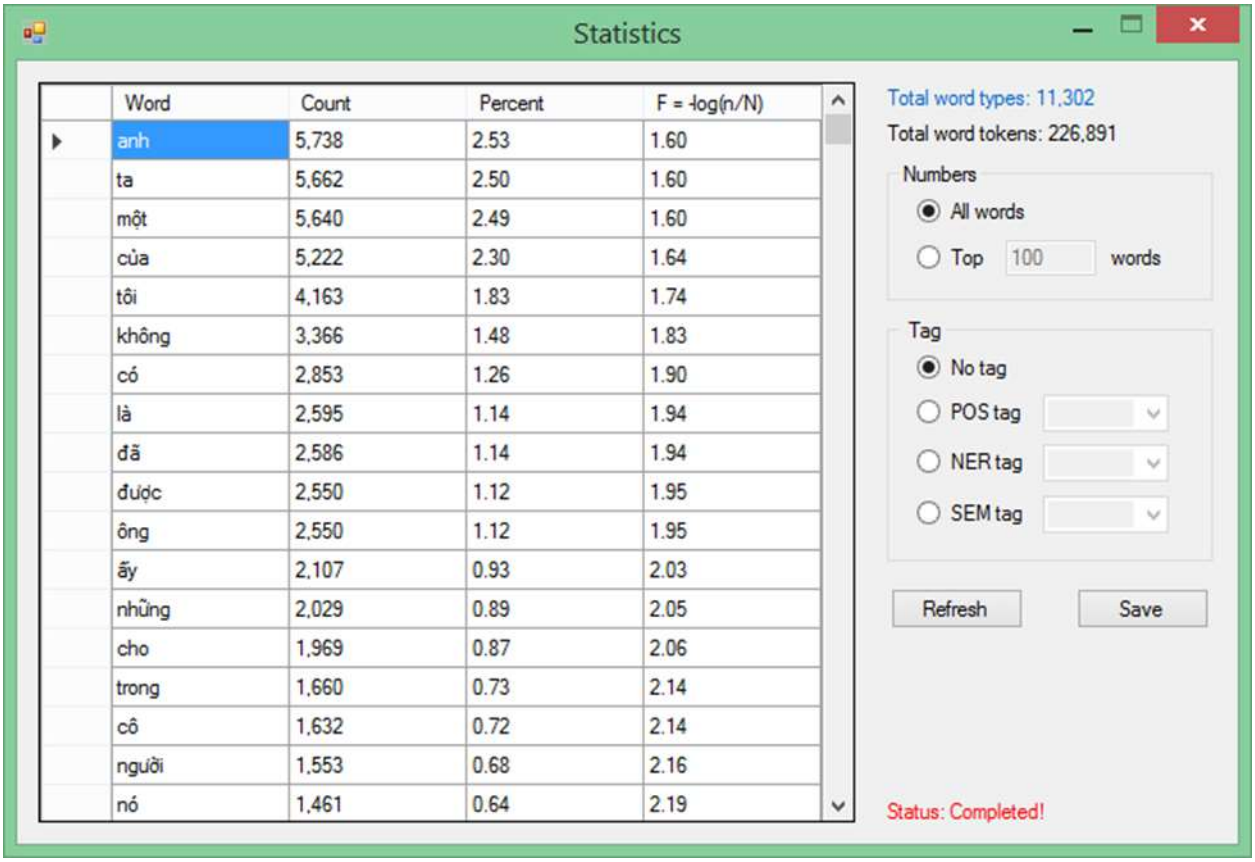
Ví dụ:

- ✓ Ta tìm từ book, dạng Morphological, không nhãn thì file lưu là: book_Morph.txt.
- ✓ Ta tìm từ book, dạng Morphological, nhãn POS là NN thì file lưu là: book_Morph_POS_nn.txt.



Hình 10: Ví dụ về tìm kiếm Word + Tags kết hợp.

4. Thống kê



Hình 11: Ví dụ về khung cửa sổ thống kê.

Tại đây ta có thể chọn ra số lượng từ có tần số thống kê lớn nhất trong khung Numbers.

Có thể thống kê theo nhãn hay không nhãn.

Có thể kết hợp cùng lúc cả 2 chức năng để thống kê.

Sau khi chọn loại chức năng, bấm nút Refresh để chương trình thống kê lại theo sự lựa chọn mới của ta.

Chương trình cho phép lưu kết quả thống kê ra file ****.csv*** để tiện lợi cho các công việc khác. File đầu ra có nội dung như khung hiển thị của chương trình, được lưu cùng thư mục chứa chương trình và có tên theo định dạng

<loại số lượng từ mà ta chọn>_<loại nhãn mà ta chọn>_<tên nhãn (nếu có)>.csv

Ví dụ:

Ta chọn là All words trong khung Numbers, No tag trong khung Tag thì file đầu ra sẽ là: ***AllWords_NoTag.csv***.

Ta chọn là Top 10 words trong khung Numbers, POS tag là NN trong khung Tag thì file đầu ra sẽ là: ***Top10Words_POS_NN.csv***.

BẢNG TRA TỪ LOẠI TIẾNG VIỆT

POS	Tag	Description
Nr	Proper noun	Proper names of people, things (Proper noun)
Nc	Classifier co-noun (loại từ)	"cái" and "con" which exist independently with single entities do not indicate nouns and can combine with numerals (Classifier co-noun)
Nu	Unit noun	Units used to measure, calculate (Unit noun)
Nt	Noun of time	Nouns imply the time (Noun of time)
Nq	Quantifier noun (numerals)	Nouns imply the quantity (Quantifier noun)
Nn	Other nouns	Single nouns, collective nouns, abstract nouns (Other nouns)
Vd	Directional verb	Verbs indicate directional actions
Ve	Exist verb	Verbs indicate the state of the entities (Exist verb)
Vc	Copula “là” verb	Verb has a special quality (Copula "là" Verb)
Vv	Other verbs	Transitive, intransitive, transformative, volition, acceptation, comparative … verbs (Other verbs)
D	Directional co-verb	Verbs stand as affixes for other verbs (Directional co-verb)
An	Ordinal number	Adjectives indicate orders, positions of entities (Ordinal number)
Aa	Other adjectives	Other adjectives (Adjective - quality, quantity)
Pd	Demonstrative pronoun	Pronouns used as affixes for another noun (Demonstratives pronoun)
Pp	Other pronouns	Other pronouns (Pronoun)
R	Adjunct	Adjuncts used to modify the meanings of verbs, adjectives or an affix other than adverb (of time, degree,…)
Cm	Major/minor preposition	Prepositions used to combine two words or two clauses of a sentence, to indicate location and time from an indicated position. (Major/ Minor preposition)

Cp	Parallel Conjunction	Conjunctions used to express syntactic relationship between two words or phrases with the same function in the sentence, or between two sentences or clauses (Parallel conjunction)
Cs	Subordinating conjunction	Conjunctions used to combine two clauses in a sentence (Subordinating conjunction)
M	Modifier word	Words used to express the speaker's attitudes (surprised, suspicious, ironic, happy…) (Modifier word)
E	Exclamation word	Words used to express emotional reactions (call, response, joy, complaint, curse, insult, …) (Exclamation word)
FW	Foreign word	Words adopted from foreign languages (Foreign word)
ON	Onomatopoeia	Words describe sounds (Onomatopoeia)
PU	Punctuation	Including all punctuations (Punctuation)
ID	Idioms	Fixed combinations of words whose meanings often cannot be interpreted simply by the words forming them.
X	Unidentified words	Words whose POS cannot be identified

BẢNG TRA NHÃN THỰC THỂ TIẾNG VIỆT

Class	Tag	Description
Abbreviation	ABB_X	Tag **ABB_X** is used for abbreviations of the normal entities and those entities listed below. E.g.: **ABB** The abbreviation of the entity does not belong to the class listed below. **ABB_DES** Abbreviations of titles **ABB_GPE** Abbreviations of geo-political entity **ABB_TRM** Abbreviations of terminology entity **ABB_ORG** Abbreviations of organizations, offices, companies **ABB_LOC** Abbreviations of places, geographic names
Title	TTL	Words indicate family relationship, social relationship or relationship in a field, profession.
Designation	DES	Position or title of a specific person.
Geo-political entity	GPE	Composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.) [2]
Person	PER	Name of a specific person or family[1], other than GPE
Organization	ORG	Names of organizations, offices or companies [1], other than GPE
Location	LOC	Names of land according to political or geographical border (city, province, country, international regions, oceans… [1] other than GPE
Date time	DTM	Time or a specific period of time [1]
Brand	BRN	Names of brands, products, trademarks.
Measurement	MEA	Measurement, quantity of things (other than money) in a standard unit.
Money	MON	Quantity of money
Percentage	PEC	Percentage of anything (other than money and standard unit)
Number	NUM	Number quantity which is not tagged as MEA or MON or PEC

Terminology	TRM	Word-combinations having special meanings depending on the contexts are used in respective specialties. They include: science, technique, military, politics, religion…