CÁC BƯỚC ĐỂ TẠO FILE XML

1. **Các file cần có để tạo file XML:**
   Bắt buộc cần có 4 file sau trong folder để tạo file XML:

- **File textgrid** (được sinh ra khi gán nhãn cho các file âm thanh). Đuôi file đúng là: **.textgrid**

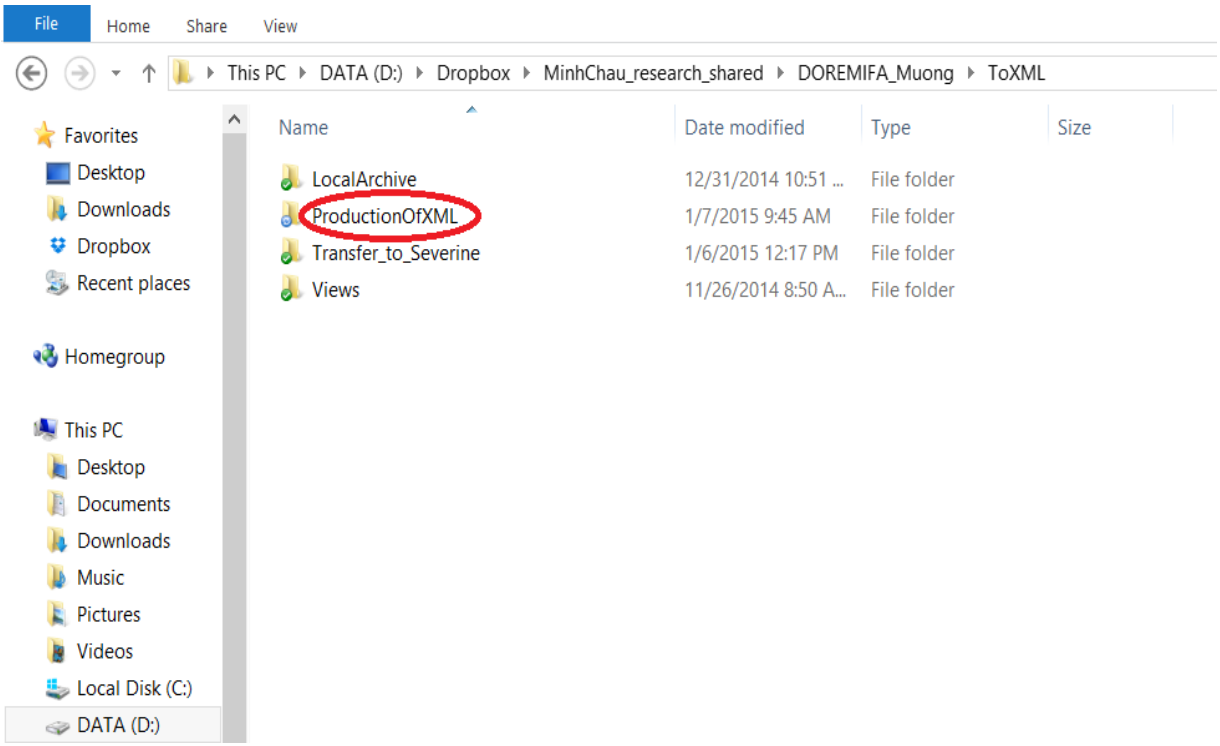
- **Wordlist** (nên dùng một bảng từ cho tất cả các file textgrid để tạo ra các file XML). Nên là bảng từ được cập nhật mới nhất. Nếu sử dụng bảng từ cũ để gán nhãn thì khi tạo các file XML nên cập nhật các thông tin mình đã bổ sung vào bảng từ mới nhất sau đó sử dụng bảng từ đó. Đuôi file là **.txt** (sau khi đã được copy sang notepad++ và save). Bảng từ bằng excel không bắt buộc phải có trong folder này.

- **Script để chạy XML (DoReMiFa_XMLGenerator.praat)**. Đã có sẵn không cần phải thay đổi gì ở file này. Đuôi file đúng là **.praat**

- **Parameters**: sau khi điền các thông tin cần thiết vào bảng biến này ta cũng copy sang notepad++ và save lại để lấy đuôi **.txt**. Trong bảng biến này, đặc biệt những dòng bôi đỏ là những dòng cần tên chính xác (cụ thể ở đây là tên của wordlist và textgrid file). Nếu không điền chính xác thì script sẽ báo lỗi ở '**praat info**' ngay sau khi chạy. Các thông tin màu xanh là các thông tin sẽ xuất hiện ở file XML sau này. Tương tự wordlist, file excel của parameters cũng không bắt buộc khi tạo ra XML file.

2. **Các bước để tạo file XML:**
   2.1. Bước 1: Tạo folder.

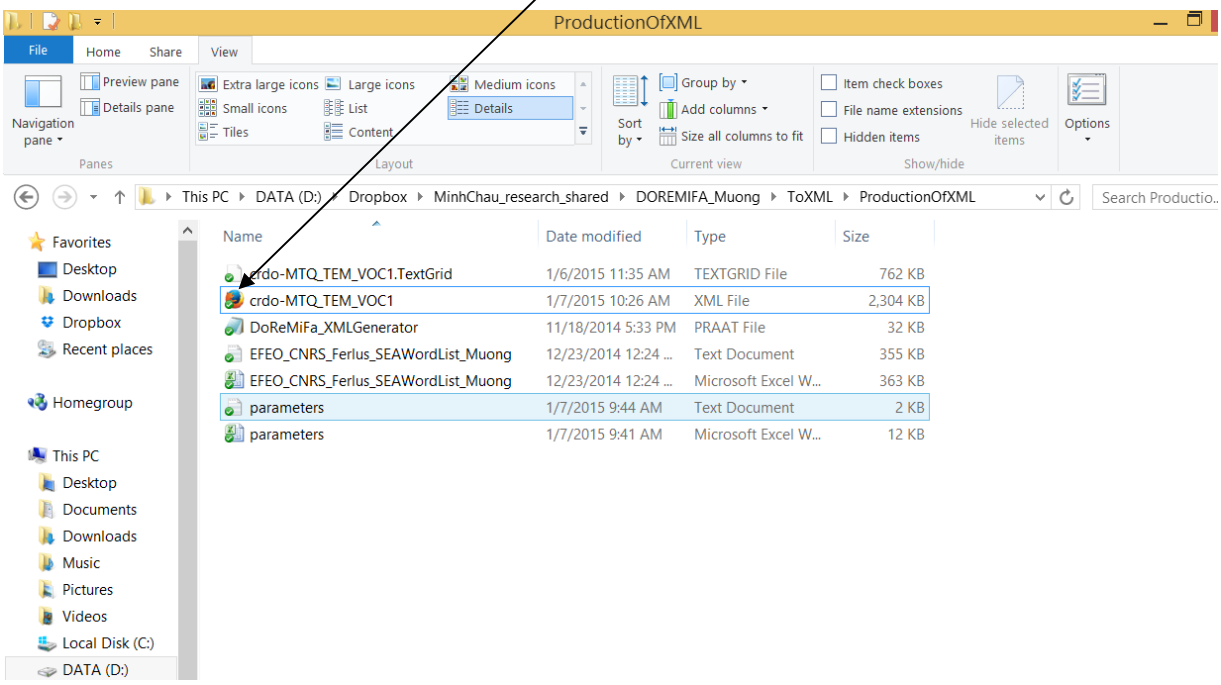
Ví dụ như folder để tạo XML của em được để trong dropbox và cố đinh trong ổ D. Trong Đó:



- "**ProductionOfXML**": là folder chính để tạo XML
- "**Views**": là folder kiểm tra sau khi đã tạo ra file XML. Folder này được cung cấp, không phải tự tạo
- "**Transfer_to_Severine**": là folder để lưu trữ các file XML đã tạo ra
- "**LocalArchive**": là folder để lưu trữ parameters, textgrid và cả wordlist (nếu sử dụng các wordlist khác nhau cho từng textgrid) sau khi đã tạo xong file XML cho file textgrid đó.

2.2. Bước 2: Đưa các thư mục cần thiết vào folder "ProductionOfXML"

Trong folder " **ProductionOfXML**" là 4 file bắt buộc phải có như trên đã nói. Cũng có thể để cả 6 file bao gồm 2 file excel để tiện sửa chữa trong quá trình sinh ra file XML. Sau khi thành công thì file XML sẽ tự sinh ra trong folder này (như ở dưới)



2.3. Chạy script (DoReMiFa_XMLGenerator.praat)

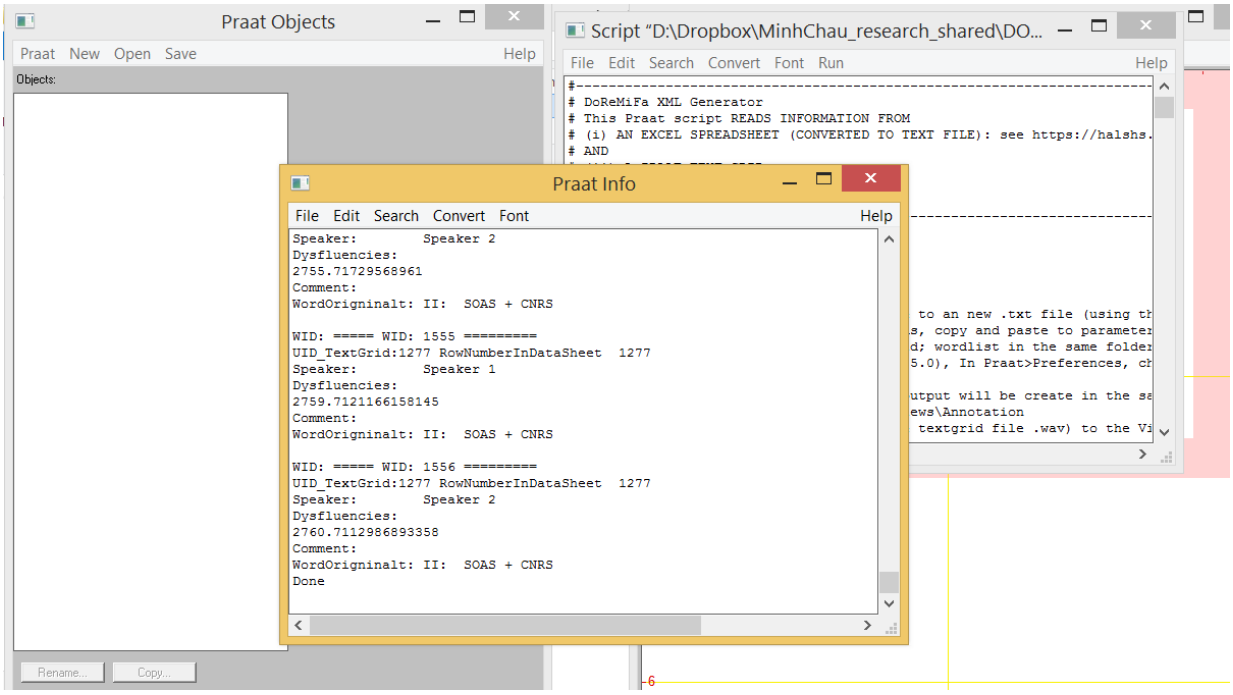
Mở script file '**DoReMiFa_XMLGenerator.praat**' bằng phần mềm praat chọn: **Praat/Open Praat Script**. Sau đó chọn Run (hoặc Ctrl+R). Quá trình chạy script có thể mất từ 5 phút đến 20 phút tùy theo dung lượng của file textgrid.

Các lỗi có thể gặp khi chạy script:

- Nếu như ngay sau khi chạy đã sinh ra '**Message**' báo lỗi thì lỗi ở đây thường là do parameters chưa tương thích, hoặc cũng có thể do lỗi ở wordlist. 'praat info' sẽ cho thông tin để biết có lỗi ở đâu. Khi đó kiểm tra lại các file đó sao cho chúng hoàn toàn tương ứng với nhau (các thông tin phải hoàn toàn giống nhau).

- Nếu như sau khi chạy được một thời gian mới sinh ra 'praat info' thì chắc chắn có lỗi ở praat. Khi ấy kiểm tra 'praat info' xem đã chạy đến UID bao nhiêu thì kiểm tra các UID ở ngay sau đó sẽ tìm ra lỗi (bằng cách sử dụng praat). Sửa và chạy lại script.

Trong khi chạy sẽ không thể xem được praat info, chỉ khi chạy thành công ta mới có thể kéo được praat infor xuống và thấy xuất hiện '**Done**' ở cuối cùng là thành công.



2.4. Kiểm tra lại

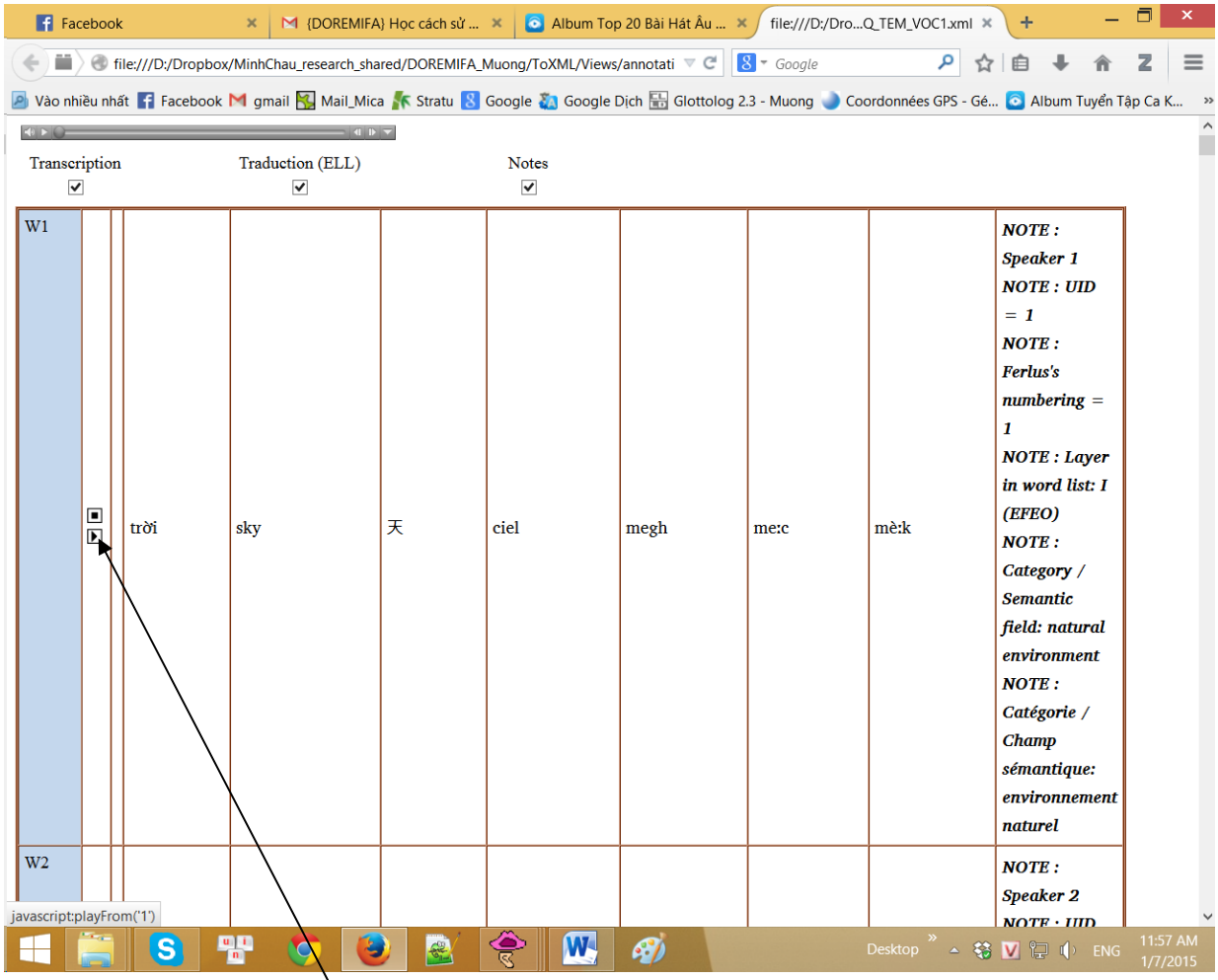
- Sau khi đã sinh ra file XML, đưa file đó vào folder **Views/annotations** (folder đã đề cập ở trên).

- Đưa file âm thanh vào **Views/audio**

- Chạy file XML bằng Mozilla Firefox (mở Mozilla Firefox → ctrl + O → chọn file XML)

- Để mở được âm thanh cần cài đặt '**QuickTime"** (có thể download đơn giản trên mạng)

- Sau khi thành công, file XML sẽ được mở trên Firefox như sau:



Nhấn vào từng từ để kiểm tra.

2.5. Lưu lại các file:

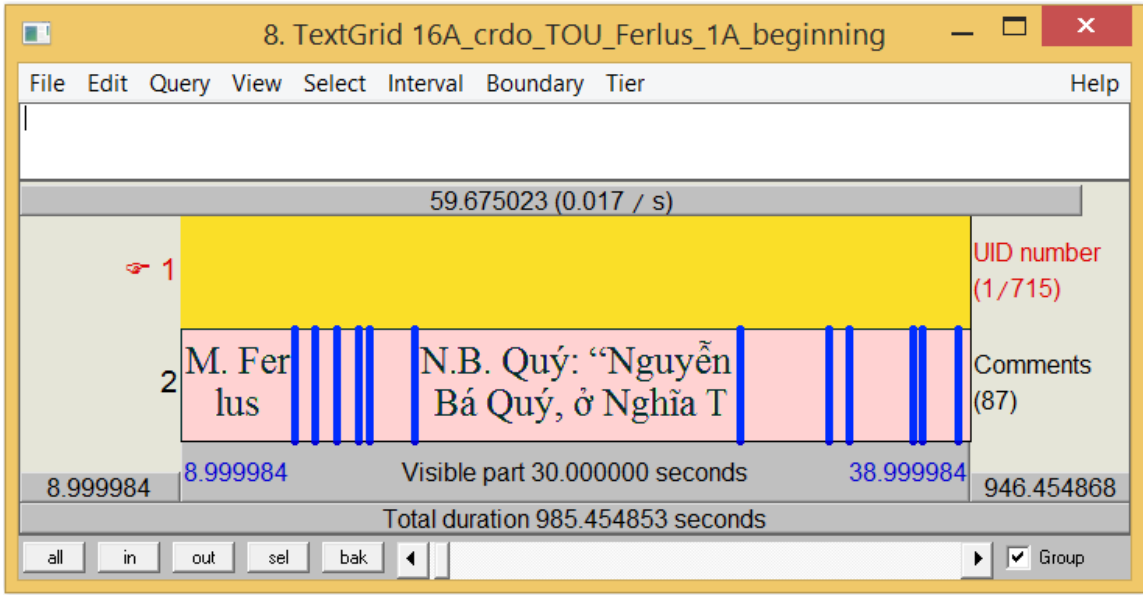
Sau khi kiểm tra và đã thành công, công đoạn cuối cùng là lưu lại các file đã sử dụng để tạo file XML và cả file XML vào các folder đã được đề cập ở trên. Trong đó:

- File XML đưa vào "**Transfer_to_Severine**"
- File textgrid, parameter (đổi tên parameter để dễ nhận diện bằng cách thêm tên của textgrid ở đầu. Ví dụ: crdo-MTQ_DANCHU_F_VOC1_parameters), và cả wordlist nếu sử dụng riêng cho mỗi textgrid, đưa vào "**LocalArchive**":
- Còn để lại các file: Script, parameters (excel), và cả wordlist (nếu dùng chung cho tất cả các tesxtgrid file) ở " **ProductionOfXML**" để tạo các XML file tiếp theo.

3. Lưu ý cho việc gán nhãn cho các file âm thanh (Annotation)

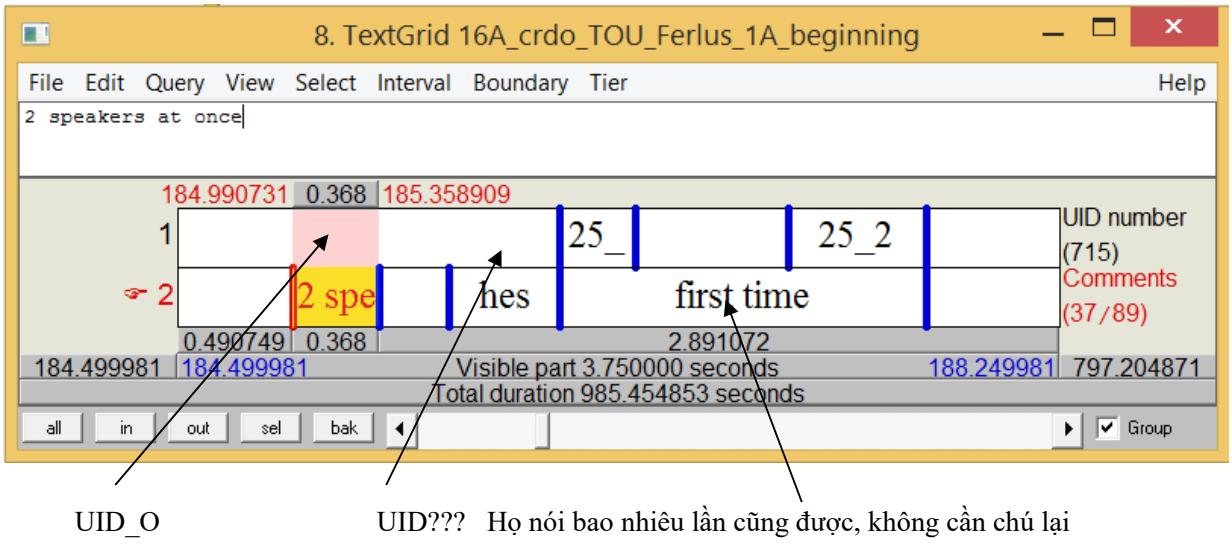
Việc này rất quan trọng trong việc tạo ra thành công 1 file XML. Có một số lưu ý khi gán nhãn như sau:

- Điều quan trọng nhất **là mỗi đoạn âm thanh (2 vạch xanh được cắt ra) phải có 1 UID tương ứng ở dòng UID number)** Nếu không có UID (ở dòng number) thì đoạn âm thanh đó không có nghĩa và sẽ gây ra lỗi. Mỗi một đoạn âm thanh (1 từ hoặc 1 cum từ) tương ứng với 1 UID. Không được cắt âm thanh mà không có UID tương ứng. Vì vậy chỉ cần cắt và chú UID cho các từ, còn các đoạn hội thoại không cần cắt và comment ở dưới.



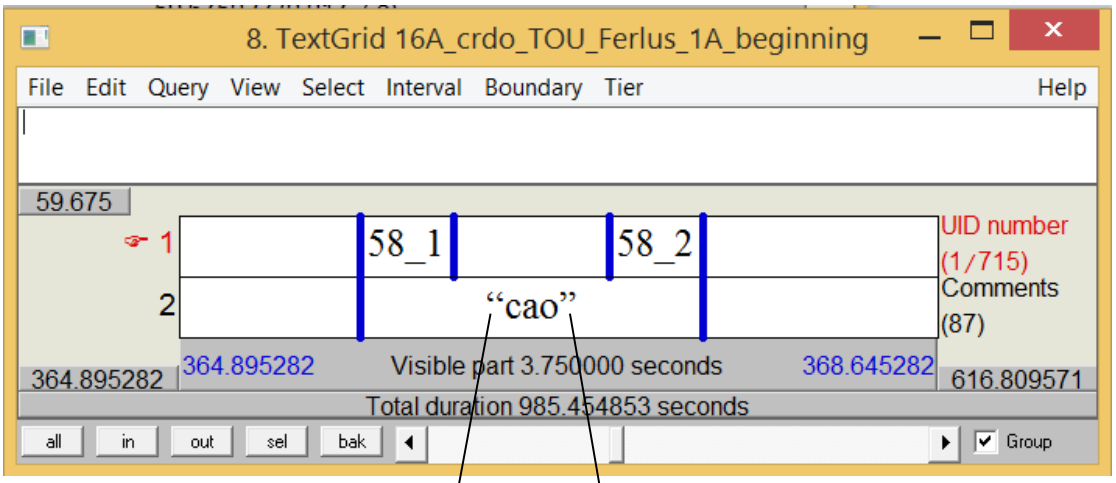
➔ Không cần thiết phải cắt những đoạn như thế này

- Tương tự, thì việc gán UID cho các phân đoạn từ là điều quan trọng nhất. Mỗi từ chắc chắn phải có 1 UID ở dòng number. Dòng comment chỉ là các ghi chú phụ cho từ đó nếu có sự không hoàn toàn tương ứng giữa từ được nói với bảng từ. **Không thể có các comments mà lại không có UID và 2 UID không thể chung 1 comment.**

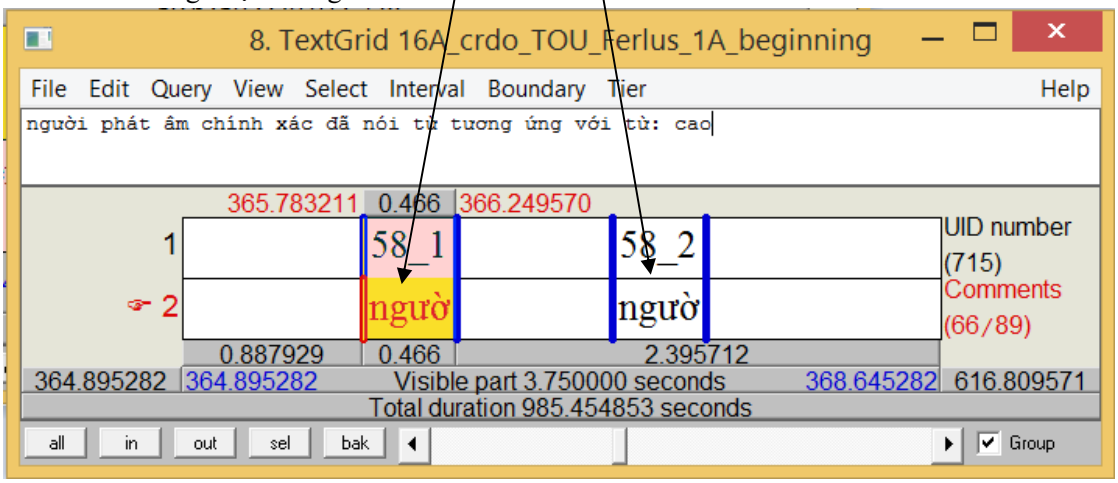


UID_O UID??? Họ nói bao nhiêu lần cũng được, không cần chú lại

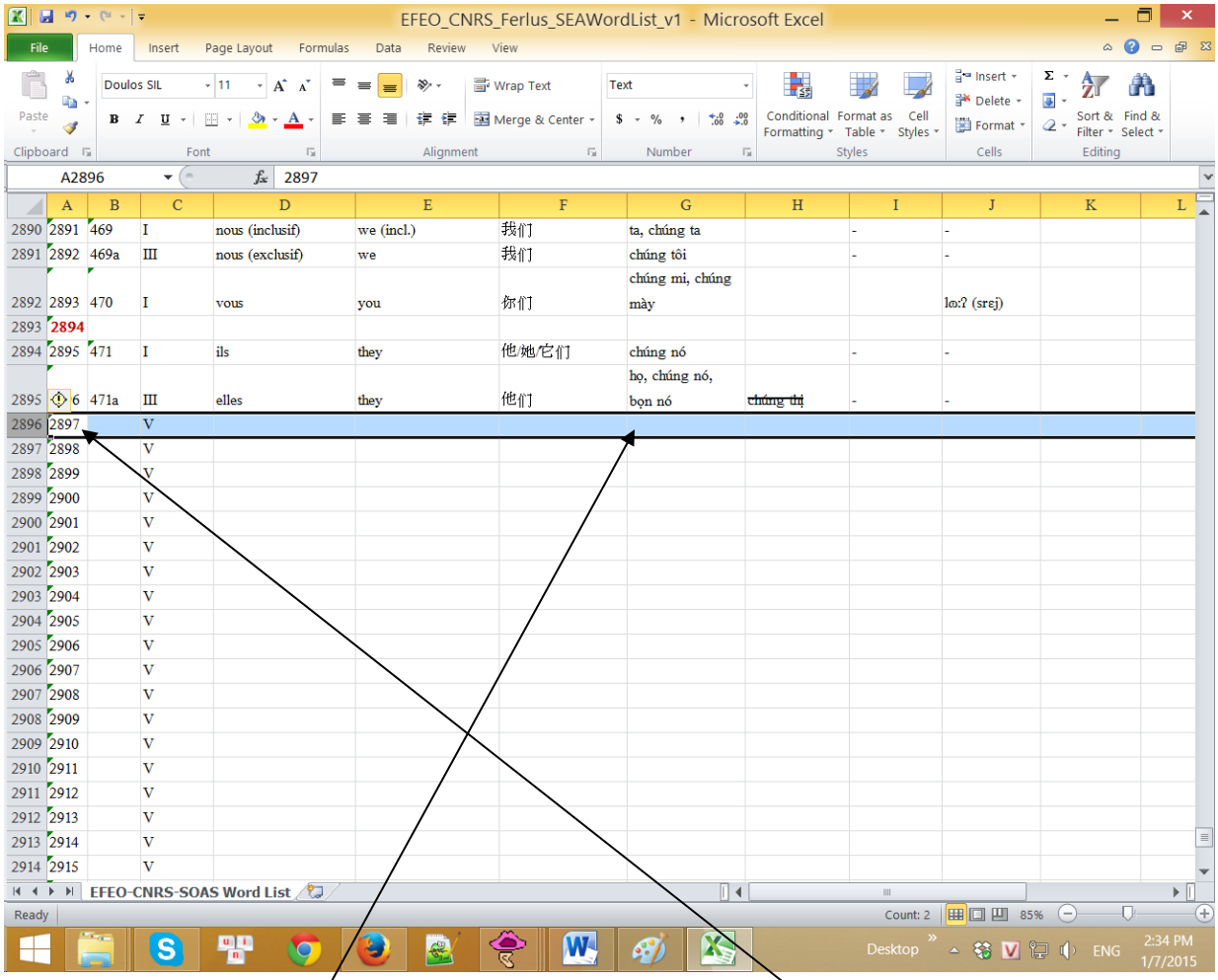
- Quan trọng là mỗi phân đoạn từ/cụm từ được gán nhãn (UID), **các UID có thể trùng nhau** (khi họ nói lại các từ đó). Nếu như từ không hoàn toàn tương ứng với từ trong bảng từ (gần nghĩa hoặc đồng nghĩa) thì ghi chú lại từ đó ở dòng comment.



➔ 2 UID không được chung 1 comment.



- Nếu như từ đó không có trong bảng từ thì thêm từ đó vào cuối cùng của wordlist và lấy UID mới đó để gán cho từ đó.



Thêm từ mới vào sau 2896 và lấy UID của từ được thêm đó để gán nhãn.

- Phần comment: có thể viết bằng tiếng Việt để giải thích sự khác biệt với wordlist hoặc có gì đó cần chú thích nên cần viết rõ ràng.

- Đặc biệt, khi gán nhãn bất kể ở dòng number hay dòng comment (và cả ở wordlist) đều không được có dấu cách và dấu xuống dòng ở đầu và ở cuối, không được sử dụng các dấu 'ngoặc đơn', "ngoặc kép" và dấu () vì đây là những dấu mặc định trong các lệnh của script.