DỮ LIỆU HỌC

- Định dạng dữ liệu học: ConLL-U format

+ Mỗi từ được thể hiện trong một dòng

+ Dòng trống đánh dấu ranh giới câu

+ Dòng chú thích bắt đầu bằng dấu #

+ Mỗi dòng gồm 10 cột, chứa các thông tin:

1.ID: Số thứ tự của từ trong câu, bắt đầu từ 1 đến hết câu

2.FORM: Từ trong câu

3.LEMMA: Gốc của dạng từ

4.UPOS: Thuộc tính chung của từ

5.XPOS: Thuộc tính cụ thể của từ

6.FEATS: Tập ngữ pháp của ngôn ngữ hay đặc trưng hình thái của từ hiện tại

7.HEAD: Thứ tự của từ trung tâm mà từ ở dòng hiện tại phụ thuộc vào

8.DEPREL: Tên quan hệ phụ thuộc giữa từ này và từ trung tâm mà nó phụ thuộc

9.DEPS: Biểu đồ phụ thuộc nâng cao ở dạng danh sách các cặp head-deprel

10.MISC: Chú thích khác

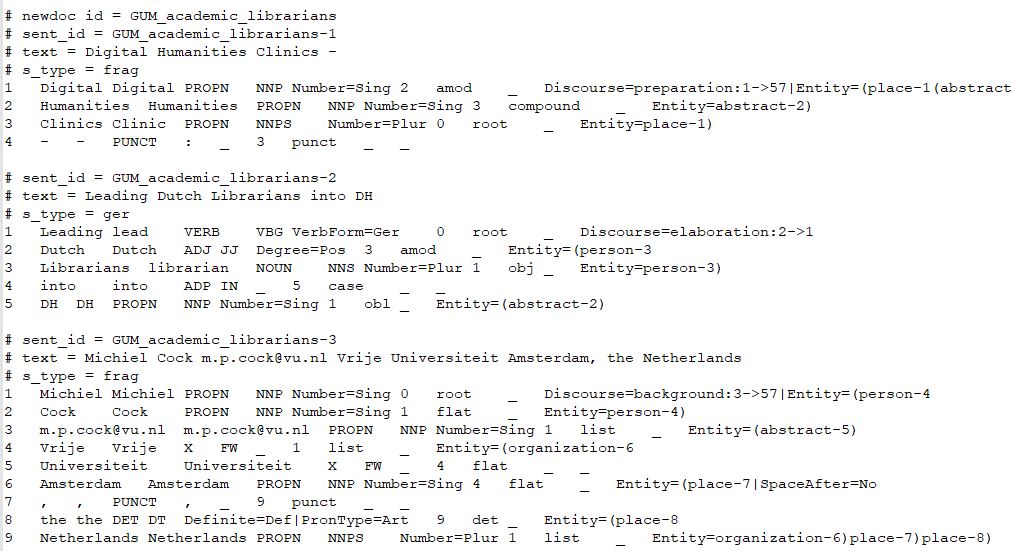
- Trong bài toán phân tích cú pháp phụ thuộc, chúng ta chỉ cần quan tâm đến các trường ID,FORM,UPOS,XPOS,HEAD,DEPREL

- HEAD,DEPREL là những thuộc tính xuất hiện sau khi phân tích cú pháp phụ thuộc

- Dấu gạch dưới biểu thị giá trị null ngoại trừ FORM,LEMMA

- FEATS,DEPS,MISC không xuất hiện có thể để giá trị null

- Ví dụ về dữ liệu học dạng CONLL\_U format:



Theo như ảnh trên, câu “Leading Dutch Librarians into DH” sẽ được phân tích cú pháp phụ thuộc ra như sau:

