



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

ERC-Project | Universität Hamburg

COBHUNI

Contemporary Bioethics and the
History of the Unborn in Islam



POS tagging and conversion of multilingual texts into Annis

Alicia González Martínez

<http://gitlab.com/kabikaj/freeling2annis>

Outline

1 Presentation

2 Motivation

3 Workflow

POS tagging: Freeling

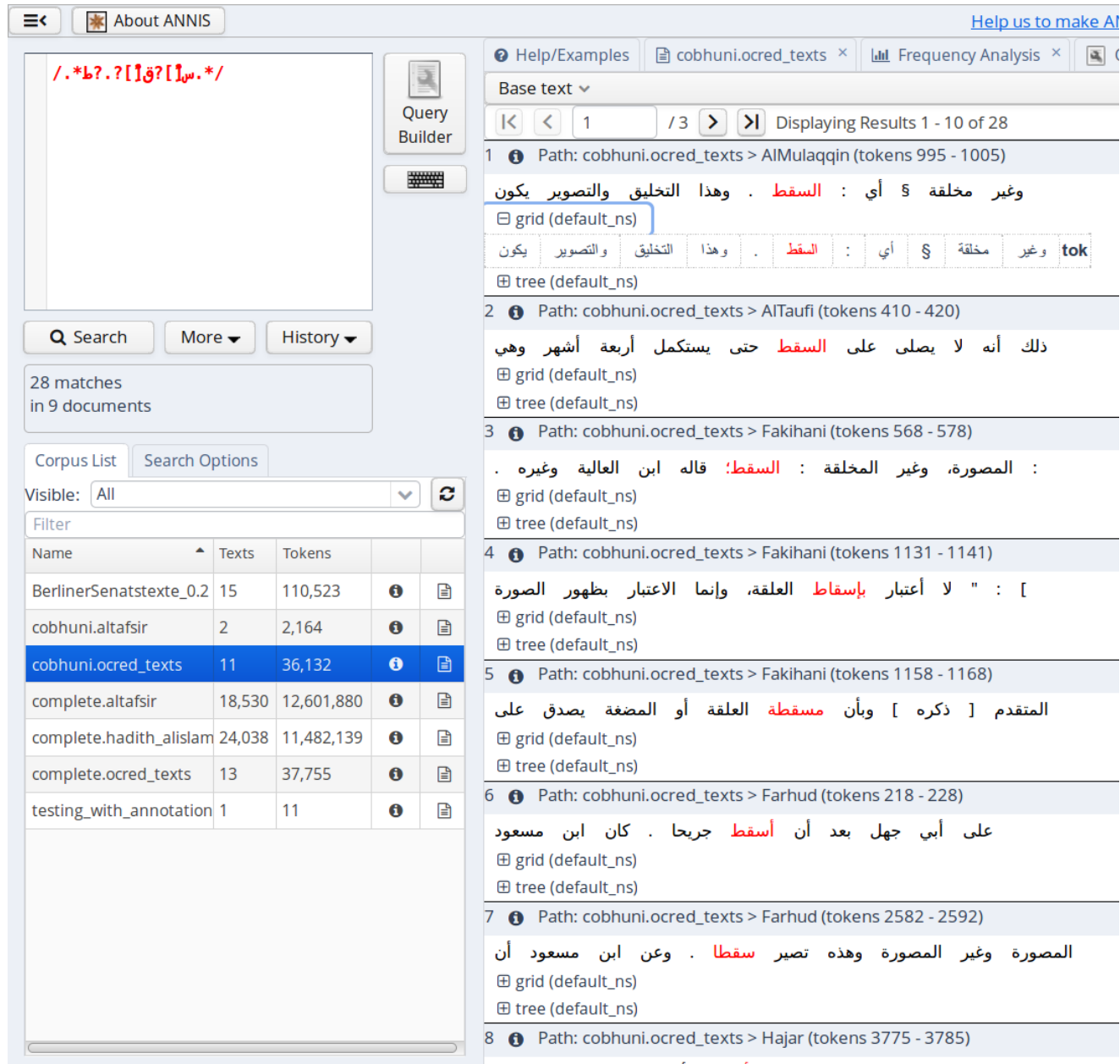
Conversion: Pepper

Visualization: Annis

<http://gitlab.com/kabikaj/freeling2annis>

Motivation

إِنَّمَا الْأَعْمَالُ بِالنِّيَّةِ وَإِنَّمَا
لِأَمْرِي مَا تَوَى فَمِنْ
كَأَنَّهُ هَجَرْتُهُ إِلَى اللَّهِ
وَرَسُولِهِ فَهَجَرْتُهُ إِلَى
اللَّهِ وَرَسُولِهِ وَمَنْ كَاتَبَ
هَجَرْتُهُ لِدُنْيَا يُصِيبَهَا أَوْ
أَمْرًا يَتَرَوَّجُهَا فَهَجَرْتُهُ
إِلَى مَا هَاجَرَ إِلَيْهِ



The screenshot shows the ANNIS search interface. The search query is `/.*ط?.?[!ق!؟[!س.*]/`. The results show 28 matches in 9 documents. The corpus list includes:

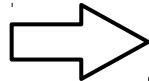
Name	Texts	Tokens
BerlinerSenatstexte_0.2	15	110,523
cobhuni.altafsir	2	2,164
cobhuni.ocrd_texts	11	36,132
complete.altafsir	18,530	12,601,880
complete.hadith_alislam	24,038	11,482,139
complete.ocrd_texts	13	37,755
testing_with_annotation	1	11

The search results on the right show the following text snippets:

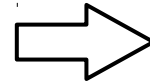
- 1 Path: cobhuni.ocrd_texts > AIMulaqqin (tokens 995 - 1005)
وغير مخلقة § أي : السقط . وهذا التخليق والتصوير يكون
- 2 Path: cobhuni.ocrd_texts > AITaufi (tokens 410 - 420)
ذلك أنه لا يصلى على السقط حتى يستكمل أربعة أشهر وهي
- 3 Path: cobhuni.ocrd_texts > Fakihani (tokens 568 - 578)
المصورة، وغير المخلقة : السقط؛ قاله ابن العالية وغيره .
- 4 Path: cobhuni.ocrd_texts > Fakihani (tokens 1131 - 1141)
[: لا أعتبر بإسقاط العلقه، وإنما الاعتبار بظهور الصورة
- 5 Path: cobhuni.ocrd_texts > Fakihani (tokens 1158 - 1168)
المتقدم [ذكره] وبأن مسقطه العلقه أو المصغرة يصدق على
- 6 Path: cobhuni.ocrd_texts > Farhud (tokens 218 - 228)
على أبي جهل بعد أن أسقط جريحا . كان ابن مسعود
- 7 Path: cobhuni.ocrd_texts > Farhud (tokens 2582 - 2592)
المصورة وغير المصورة وهذه تصير سقطا . وعن ابن مسعود أن
- 8 Path: cobhuni.ocrd_texts > Hajar (tokens 3775 - 3785)

Workflow

Perder la mirada,
distráidamente,
perderla y que nunca la
vuelva a encontrar:
y, figura erguida, entre
cielo y playa,
sentirme el olvido perenne
del mar.

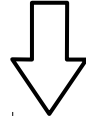


FreeLing 4.0

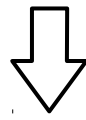


form lemma POS

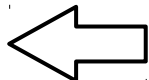
```
....  
cielo cielo NCMS000 1  
y y CC 0.999989  
playa playa NCFS000 1  
, , Fc 1  
sentir sentir VMN0000 1  
me me PP1CS00 1  
el el DA0MS0 1  
olvido olvido NCMS000 0.989726  
perenne perenne AQ0CS00 1  
de de SP 1  
el el DA0MS0 1  
mar. marzo NCMS000 1
```



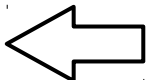
freeling2xml.py



```
<?xml version="1.0" encoding="utf-8"?>  
<a><b/>  
<content>  
<t POS="VMN0000" lemma="perder">Perder</t>  
<t POS="DA0FS0" lemma="el">la</t>  
<t POS="NCFS000" lemma="mirada">mirada</t>  
<t POS="Fc" lemma=",">,</t>  
<t POS="RG" lemma="distráidamente">distráidamente</t>  
<t POS="Fc" lemma=",">,</t>  
<t POS="VMN0000" lemma="perder">perder</t>  
<t POS="PP3FSA0" lemma="lo">la</t>  
<t POS="CC" lemma="y">y</t>
```



pepper



createPepperWorkflow.py

