CORPUS LINGUISTICS

What is it and what can it do for me?

1st IDHN Members’ Meeting, May 24, 2019

Jakob Prange

Overview

- What is a Corpus?
- Why Corpus Linguistics?
- Where can I start?

WHAT IS A CORPUS?

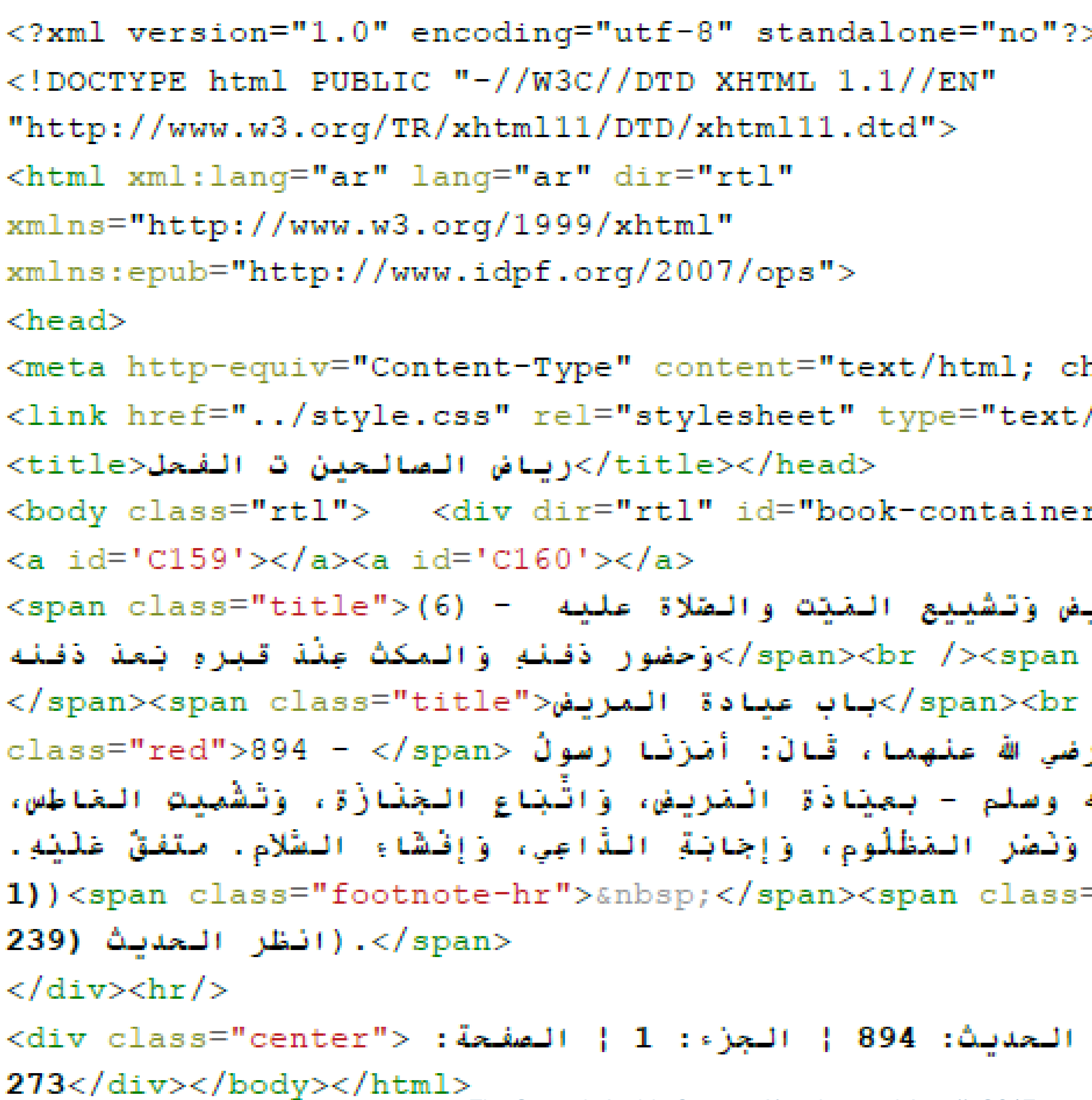
What is a corpus?

A corpus (from Latin: body) is a large collection of texts.

Often structured or annotated.

Nowadays stored electronically.

Thus easy to process with computational methods.



```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
"http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xml:lang="ar" lang="ar" dir="rtl"
xmlns="http://www.w3.org/1999/xhtml"
xmlns:epub="http://www.idpf.org/2007/ops">
<head>
<meta http-equiv="Content-Type" content="text/html; ch
<link href="../style.css" rel="stylesheet" type="text/
<title>رياض الصالحين ت الفحل</title></head>
<body class="rtl">   <div dir="rtl" id="book-containe
<a id='C159'></a><a id='C160'></a>
<span class="title">(6) -  ض وتشييع الميت والصلاة عليه
وحضور دفنه والمكث عند قبره بعد دفنه</span><br /><span
</span><span class="title">باب عيادة المريض</span><br
class="red">894 - </span> رضي الله عنهما، قال: أمرنا رسولُ
ه وسلم - بعيادة المريض، واتّباع الجنازة، وتشميت العاطس،
ونصر المظلوم، وإجابة الدّاعي، وإفشاء السّلام. متفق عليه.
1))<span class="footnote-hr">&nbsp;</span><span class=
239) انظر الحديث).</span>
</div><hr/>
<div class="center"> الحديث: 894 | الجزء: 1 | الصفحة:
273</div></body></html>
```

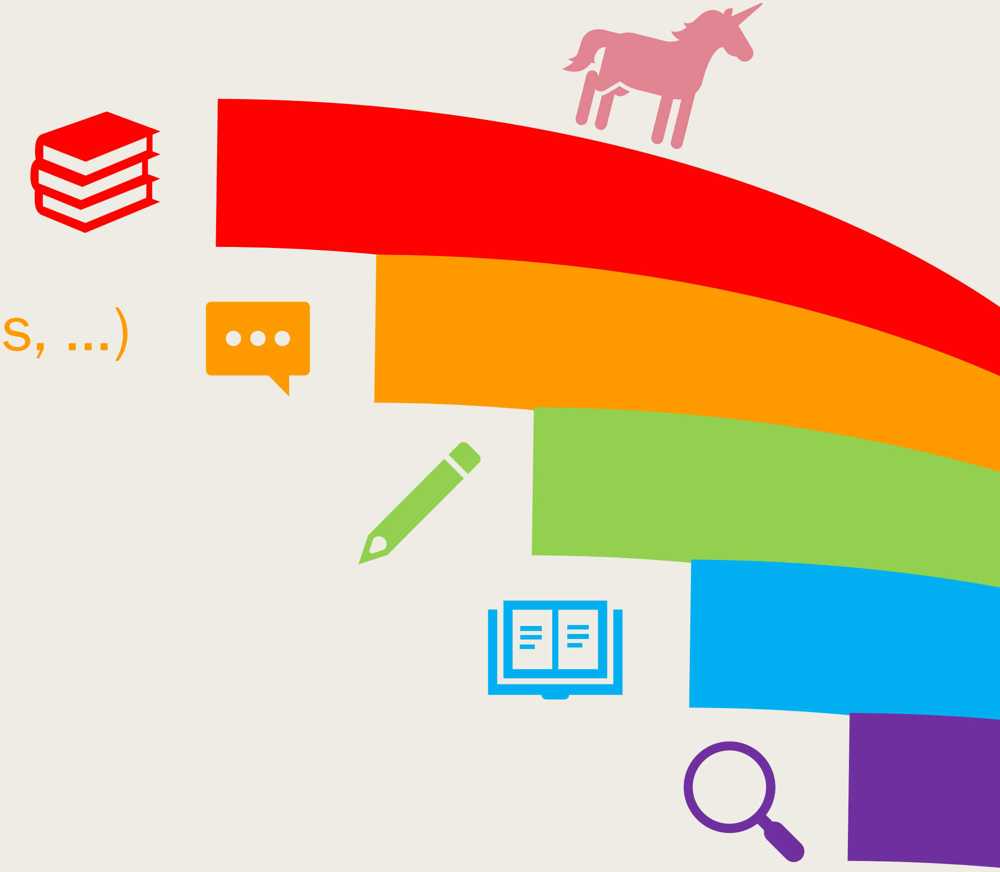
The Sunnah Arabic Corpus, Alosaimy and Atwell, 2017

When working with corpora…

… you **have access** to many things:

- Data (e.g., from Quran, Hadiths)
- Metadata (title, author, paragraphs or verses, …)
- Annotations (linguistic, interpretative, …)

- Translations (in “parallel corpora”)
- Software (search engines, APIs)



When working with corpora…

… you have to deal with many things!

- Data **formats** (raw text, XML, TEI, …)
- **Encodings** (alphabets, direction of writing)
- **Licensing** (open source or paid)

When *building* corpora...

... you have to deal with **even more** things!

- Data formats (raw text, XML, TEI, ...)
- Encodings (alphabets, direction of writing)
- Licensing (open source or paid)
- **Where and how** to get the data and annotations
- Data **filtering** and **validation**
- **Representativeness** and **balance** (across languages, genres, words)

WHY CORPUS LINGUISTICS?

Why Corpus Linguistics?

- Lots of data in one place

- **Efficient** *statistical and qualitative* analyses
  (easier if appropriate software is available)

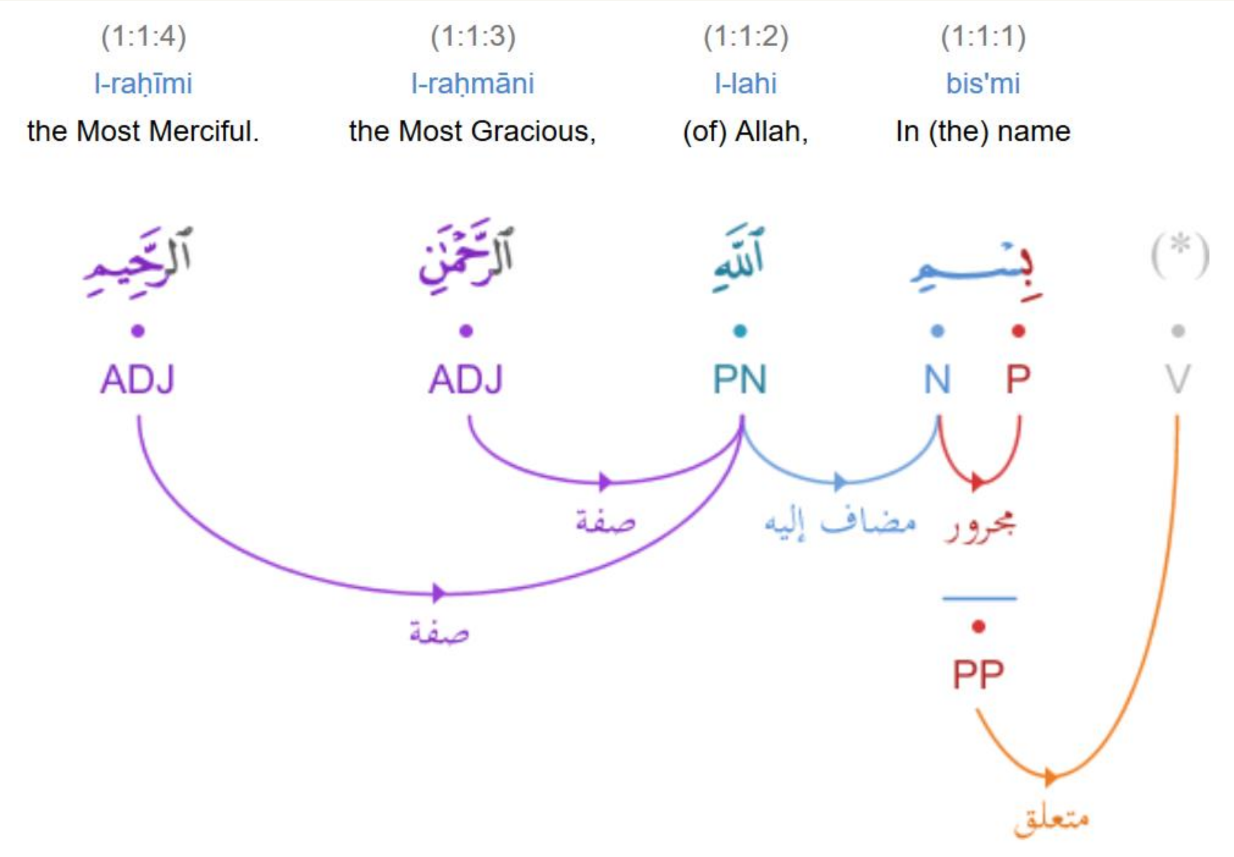
- Can be used to create or test software,
  for future analyses or non-academic users

- Needed for training machine learning systems
  (e.g., automatic morphological or syntactic analyzers)

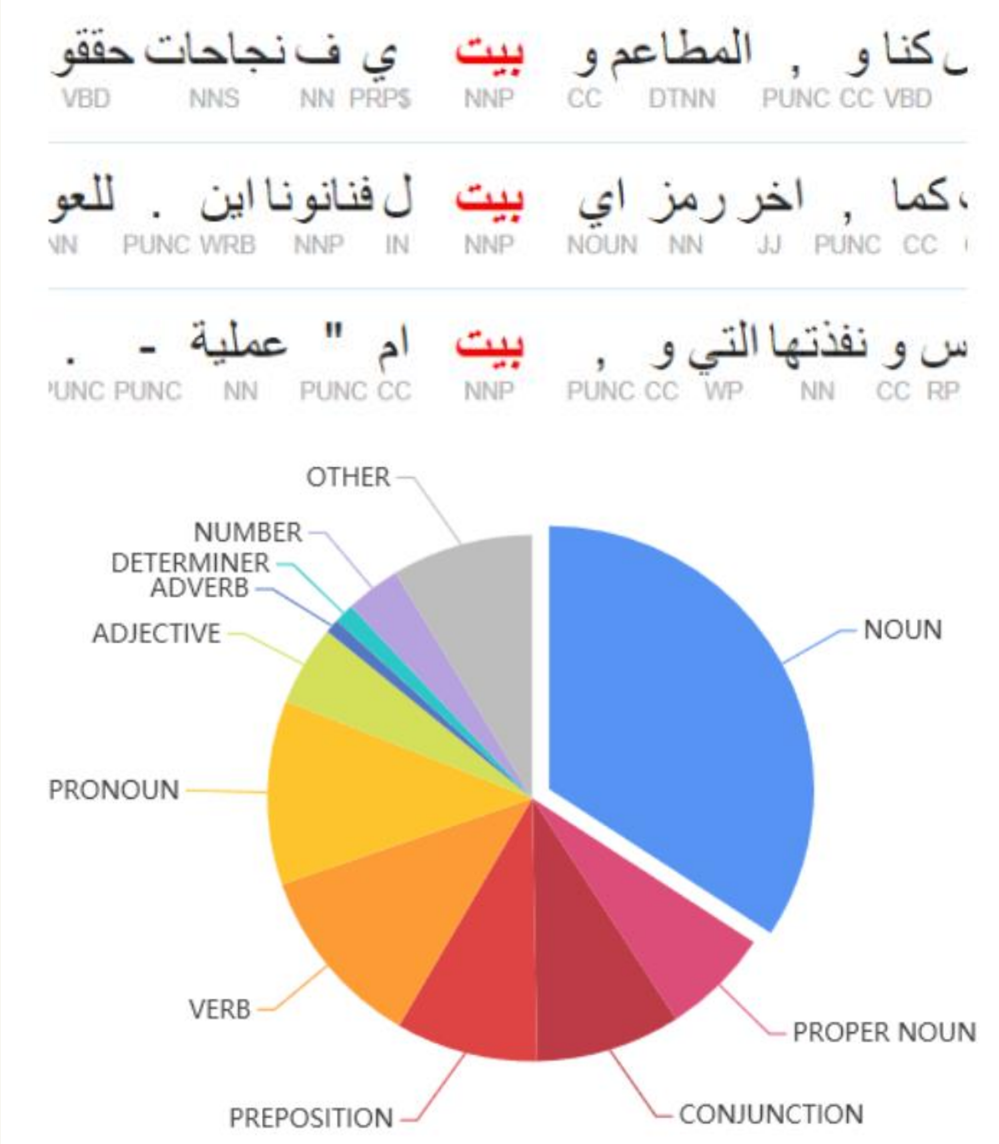
WHERE CAN I START?

Specific Resources

- The Quranic Arabic Corpus
  - *Syntactic treebank, morphological annotation, semantic ontology, English translation*
- Sunnah Arabic Corpus
  - *Follows QAC in its standards*
- 12er Hadith Corpus
  - *Translation and comments, but no structural annotations*



More Resources



- List of Arabic corpora
- Sketch Engine
  - *Collocations, thesaurus, frequency lists, examples in context*
  - *Need paid account*

Publishers

- Linguistic Data Consortium (LDC)
- European Language Resources Association (ELRA)