-HAT-

corpus 2020

by H. Sayoud / EDT

Description of the HAT Corpus

Our textual dataset is composed of 100 groups of Arabic texts that are extracted from 100 different Arabic books. The books were written by 100 different authors and each group contains 3 different texts that are written by the same author, which means that each group belongs to only one author. This set of 300 text documents has been collected in 2019 from "*Alwaraq digital library*" (*www.alwaraq.net*).

We caled this corpus "HAT" (*i.e. Hundred of Arabic Travelers*). Furthermore, the HAT corpus could represent a reference dataset for author style analysis in Arabic, which could be used by NLP researchers for a purpose of comparative evaluation. More details on HAT can be found at *www.univ-edt.com*.

For concreteness, here is a piece of text belonging to Author #92 (*figure 1*).

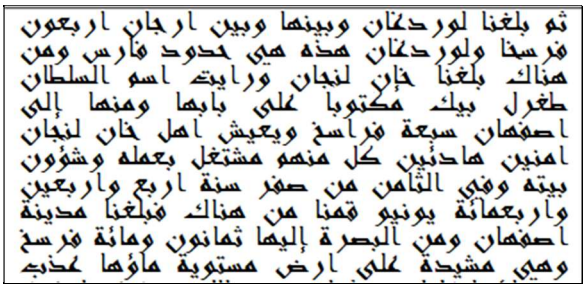
ثم بلغنا لورد كان وبينها وبين ارجان اربعون
فرسخا ولورد كان هذه هي حدود فارس ومن
هناك بلغنا خان لنجان ورايت اسم السلطان
طغرل بيك مكتوبا على بابها ومنها إلى
اصفهان سبعة فراسخ ويعيش اهل خان لنجان
امنين هادئين كل منهم مشتغل بعمله وشؤون
بيته وفي الثامن من صفر سنة اربع واربعين
واربعمائة يونيو قمنا من هناك فبلغنا مدينة
اصفهان ومن البصرة إليها ثمانون ومائة فرسخ
وهي مشيدة على ارض مستوية ماؤها عذب

Fig. 1. Example of Arabic text belonging to Author #92 (*N. Khasru*)

Fig. 2. Ancient portrait and sheet of paper containing a text of Author #92 (*N. Khasru*).

List of the authors and their corresponding topics

The different authors are listed in table 1.

Table 1. List of Authors.

Reference Number	Author Name	Author name in Arabic	Reference Number	Author Name	Author name in Arabic
1	Ibrahim Basha	إبراهيم رفعت باشا	51	Sara Sirayt	ساراسيرايت
2	Ibrahim Almazini	إبراهيم عبد القادر المازني	52	Said ben Ali	سعيد بن علي المغيري
3	Ibn Almujawer	ابن المجاور	53	Saloum Addahdah	سلوم الحدحاح
4	Ibn Batuta	ابن بطوطة	54	Suleiman bnu Sayam	سليمان بن صيام
5	Ibn Djubair	ابن جبير	55	Sadeq Bacha	صادق باشا المؤيد العظم
6	Ibn Fathlane	ابن فضلان	56	Taleb Muchtraq	طالب مشتاق
7	Ahmed Atelemsani	أحمد بن مطال التلمساني	57	AbdAlHalim Almisri	عبد الحليم المصري
8	Ahmed Albaqri	أحمد ماهر محمود البقري	58	AbdAlGhani Alnabulsi	عبد الغني بن إسماعيل النابلسي
9	Ahmed Hassaneen	أحمد محمد حسنين	59	AbdAllatif Albaghdadi	عبد اللطيف البغدادي
10	Ikhraj Arradhi	إخراج ابتهاج عمر طاهر الراضي	60	AbdAllah Afandi	عبد الله أفندي إلياس رعد
11	Ismaeel Mustapha	إسماعيل محمد مصطفى	61	AbdAllah Alkandary	عبد الله بن محمد بن سالم باكثير الكندي
12	Alab Malon	الأب ألكسيس مالون	62	AbdAllah Heshima	عبد الله حشيمه
13	Alab Rene	الأب رينه موتريد اليسوعي	63	AbdAlMuhsin Albarkaty	عبد المحسن البركاتي
14	Alab Luis	الأب لويس جلابرت اليسوعي	64	Abdalmasseeh Antaky	عبد المسيح أنطاكي
15	Alab Michel	الأب ميشال جوليان اليسوعي	65	Azam AbdAlWahab	عزام عبد الوهاب
16	Alab Henry	الأب هنري لامنس اليسوعي	66	Ali Ibrahim	علي إبراهيم كردي
17	Ameer Mohamed Ali	الأمير محمد علي باشا	67	Ali Aldjerjawy	علي أحمد الجرجاوي
18	Albalwee	البلوي	68	Fouad Ghosn	فؤاد غصن
19	Alhadj Ahmed Kamal	الحاج أحمد كمال	69	Cartsten Nebur	كارستن نيبور
20	Alkhouri Batras	الخوري بطرس العنداري	70	Clodis Jems	كلوديس جيمس ريج
21	Almudhfir Almusili	الرحالة العربي المظفر الموصلي	71	Alexander Kinglake	كينغليك ، نقلها إلى العربية محمود العابدي
22	Seddeek Hassan Khan	السيد صديق حسن خان	72	Abu AbdAllah Alabdary	لأبي عبد الله محمد بن محمد العبدري الحيحي
23	Muhsen Abutabeekh	السيد محسن أبو طبيخ	73	Lessan Eddin bnu Alkhatib	لسان الدين ابن الخطيب
24	Muhsen Alameen	السيد محسن الأمين	74	Alab Luis Ranz	للأب لويس رنز
25	Muhamed ibn Assayed	السيد محمد ابن السيد احمد الحسيني	75	Leonhard Rauwolf	للرحالة الهولندي الدكتور ليونهارت راوولفه
26	Sheikh Muhamed Alqayati	الشيخ محمد عبد الجواد القاياتي	76	Mar Athanasius	مار اثناسيوس اغناطيوس نوري
27	Alqas Ishaq	القس اسحق أرملة	77	Muheb Eddine Alhamwy	محب الدين الحموي
28	Alqas Antonio	القس انطونيوس شبني اللبناني	78	Muhamed Alhajwy	محمد الحجوي
29	Alqas Suleiman	القس سليمان صائغ الكلداني الموصلي	79	Muhamed Alhur	محمد الحر
30	Othman Multadha	المطبعة الأميرية بالقاهرة / عثمان مرتضى	80	Muhamed Alhashemy	محمد الهاشمي
31	Alyussy	اليوسي	81	Muhamed Amine	محمد أمين فكري بك
32	Benyamin Alandalussy	بنايمين يونة التطيلي النباري الأندلسي	82	Muhamed Aumissawy	محمد بن عبد الله الحسيني الموسوي
33	Imad Raouf	تحقيق د. عماد عبد السلام رؤوف	83	Muhamed Thabet	محمد ثابت
34	Seyar Aldjameel	ترجمة د. سيّار الجميل	84	Muhamed Rashed Reda	محمد رشيد رضا
35	Batras Haddad	ترجمة وتعليق الأب د. بطرس حداد	85	Muhamed Saoud	محمد سعود العوري
36	Mustapha Juda	تعريب مصطفى محمد جودة	86	Muhamed Kurd	محمد كرد علي
37	Djerji Serkis	جرجي بن يوسف اليان سركيس	87	Muhamed Labeeb	محمد لبيب البتنوني
38	Djerjee Zidan	جرجي زيدان	88	Mahieddine Reda	محي الدين رضا
39	John Luis	جون لويس بوركهارت	89	Mustapha Muhamed	مصطفى محمد
40	HarethnYussuf	حارث يوسف ثنيمة	90	Naji Jawed	ناجي جواد
41	Habeeb Afendy	حبيب أفندي فريحه البشعلاني	91	Nasser Djarjees	ناصر جرجيس
42	Hassan Muhamed	حسن محمد جوهر	92	Nasser Khasru	ناصر خسرو
43	Hassan Fawzi	حسين فوزي	93	Nazih Muayed	نزيه مؤيد العظم
44	Alkhouri Ibfahim	حضرة الخوري إبراهيم حرفوش	94	Newab Hamid	نواب حميد يار جونك بهادر
45	Alkhouri Luis	حضرة الخوري لويس ملحة	95	Nawal Assaadawy	نوال السعداوي
46	Siraz Dubler	حققها سيزار دوبلر	96	Henry Tristram	هـ. ب. تريسترام
47	Hamad Aljasser	حمد الجاسر	97	Hor Diver	هور ديفر
48	Sherif Hetata	د. شريف حتاتة	98	Yahia bnu Muhassen	يحي بن أبي الحسن أحمد المعروف بابن محاسن
49	Refaa Attahtawi	رفاعة الطهطاوي	99	Yussof Aboud	يوسف عبود
50	Arrabie ben Suleyman	رواية تلميذه الربيع بن سليمان الجيزي	100	Youssof Kamal	يوسف كمال

The HAT text documents have a relatively short/medium size: the average text length is about 1100 words per document and there are 3 documents per author, which corresponds to 300 documents in the entire corpus. This situation involves severe experimental conditions, since it has been shown, in previous research works (Eder 2010;

Signoriello 2005), that the minimum number of words per text should be at least 2500 words to get good attribution performances. In this investigation, the use of medium-short texts is interesting in order to evaluate the different classifiers with small documents in Arabic. In fact, when short texts are used, the AA performances decrease and it becomes difficult to make an efficient identification.