Annotation Guidelines for Spatial Named Entities in Arabic

This work is used as part of the article:

Alrahabi Motasem, Brando Carmen, Alkhalil Muhamed, Dichy Joseph (2020), "Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie d'entités nommées de lieu", Revue Humanités numériques, Humanistica, 2020.

Definition:

Named Entity Recognition is a subtask of NLP aiming to identify real-world entities in texts, such as names of persons, organizations, and locations, among others (Nadeau and Sekine, 2007 ; Nouvel et al., 2015). Examples of named entities:

أبو ظبي, الرئيس شيراك, الأمم المتحدة

Abu Dhabi, President Chirac, United Nations

Classes:

In this work, we are only interested in Locations (places) that have a unique referential entity and ideally coordinates. This category can be divided into multiple fine-grained types:

- **Natural and geographical places** (LOC-Nature): deserts, mountains, rivers, seas, planets, galaxies…:
  - السين ,جبل بلقان ,بحر القلزم ,صحراء الحجاز ,عطارد ...
- **Administrative regions** (LOC-Admin): demarcated by humans using imaginary boundaries: neighborhood, country, continents, borders...:
  - ليون ,مدينة أسيوط ,بلاد النمسا ,آسيا ,حدود السودان...
- **Buildings and functional constructions** (LOC-Building): town hall, shop, church, school…when they designate a place and not an organization :
  - بلدية باريس ,مدرسة أركان الحرب ,مكتب اللغات المشرقية ,أكدمية العلوم الأدبية, الجامع الأزهر ,مجلس محافظة القاهرة ,قصر لقسمبورغ ...
- **Paths and axes of traffic** (LOC-Path):line, street, avenue, tunnel…:
  - شارع القاهرة ,الجسر المعلق ,نفق الاهرام ,ساحة شمدين ,باب سان دوني...

General remarks:

1. We do not annotate NE that are not written in Arabic.

2. We do not annotate geo-political entities or organizations:
   - يقول اليونان أن..., تسعى فرنسا للسيطرة...(ولكن نأخذها عندما تدل على مكان: تقع فرنسا في اوربا...)

3. We do not annotate anaphoric expressions for a NE previously occurred in the discourse:
   - واندلعت فيها حرائق كبيرة.

4. We do not annotate the adjectives around the NE:
   - باريس الرائعة (ولكن نأخذها ان كانت من أصل المسمى مثل الشرق الأقصى, البحر الأحمر...)

5. We do not annotate the agglutinated clitics (prefixes, suffixes) such as:
   - الباء :بالأندلس ...
   - الكاف :كبلاد الصين ...
   - الفاء :فأمريكة الشمالية, فبلاد الحجاز ...
   - الواو :والبحر الأبيض, وبمدينة باريس ...
   - اللام :لبلاد آسيا, فلبلاد آسيا, ولبلاد آسيا...
   - اللام الملتصقة بألف التعريف في اسماء بعض البلدان: لليابان, للطائف...
   - التنوين :أزهراً, مسجداً...
   - أدوات الملكية: مصرنا (ونتجاهل حالياً كل المسميات المركبة التي تحتوي أداة ملكية: دولتنا المصرية)

6. We annotate the determiner الـ that precede the NE if it is not attached to another prefix:
   - الجامع الأزهر, المسجد الحرام, البلاد الإفرنجية...(ولكن لا نأخذ الباء أوغيرها كما في: بالمسجد الحرام)

7. We do not annotate the spatial relations, the whole to part relations, the indoor locations, the partitive constructions…:
   - داخل الحرم الشريف, خارج المدرسة الوطنية للعلوم ...
   - تجمعوا في مدرج كلية الآداب.
   - جنوب فرنسا, شرق سوريا... (ولكن نأخذ بلداناً مثل جنوب أفريقيا...)

- جانب ساحة النصر ,بالقرب من جامعة فاس...
- جزء من حلب تدمر بالكامل.
- كثير من / بعض المدن الفرنسية
- أصغر مدينة فرنسية
- ...

8. We annotate the generic part of a place name:

- شارع بغداد ,ساحة الشهداء ,كرسي البرتوغال ,تخت الدولة العلية ,بوغاز جبل طارق...(ولكن لا نأخذ هذه الكلمات عندما تكون بعيدة عن المسميات, مثل المارستان المسمى أوتيل ديو...)

9. We do not annotate expressions or common names that do not include a proper name:

- مشيت في هذا الشارع, أحب بلدي, تصلي في الكنيسة, نسبح في النهر ...

10. We annotate the widest portion of a NE, ignoring nested structures:

- رئيس جامعة نيويورك أبوظبي.

11. We annotate the NE aliases:

- مملكة المصريين ,ممالك الإسلام ,أرض الروم ,امبراطورية قيصر ,مدينة الأنوار ,كرسي ملك الاندلس ,بلاد الفرنسيس...(ولكن لا نأخذ مثل هذه الكلمات عندما تكون متباعدة في السياق, كما في المثال: بلاد الفرس, واليمن, والروم, والفرنجة)

12. We annotate all differents transcriptions of the same location:

- أفغانستان ,أفغنستان /...أمريكيا ,أمريقة ,أمريكا...

13. We do not annotate imaginary places.

14. We do not annotate places on a small scale (bedroom, office, living room...) unless they are accompanied by geographical details (*the oval office of the White House*).