Guide d'annotation des entités nommées au sein d'un corpus persan

Ronan Grandadam, Kévin Deturck, Damien Nouvel

Guide d'annotation des entités nommées au sein d'un corpus persan

Le projet VITAL (Valorisation de l'Innovation dans le Traitement Automatique des Langues) est le fruit d'une collaboration entre le laboratoire ERTIM de l'INALCO, équipe de recherche en ingénierie multilingue et en Traitement Automatique des Langues (TAL), et l'entreprise Kairntech, éditeur de logiciels. Le projet VITAL vise à développer des outils performants d'assistance à la production de données annotées et à permettre à ces outils de mieux supporter certaines langues moyennement ou peu dotées (kurde, ourdou, hindi, persan, mandarin, cantonais, arabe standard, arabe tunisien). Dans le cadre du projet, différentes campagnes d'annotation de corpus ont eu lieu, dont une sur un corpus persan.

Aujourd'hui, la production de données annotées est au cœur d'enjeux importants, dans la mesure où les systèmes neuronaux d'intelligence artificielle ne fonctionnent pas sans données, en partie issues d'annotations réalisées manuellement par des locuteur·ices humain·es.

Entre autres types d'annotations, celle des entités nommées permet de faire progresser les outils de reconnaissance automatique des éléments informationnels au sein de textes[1]. Ces outils peuvent être utiles, entre autres, dans le cadre du développement des moteurs de recherche ou des moteurs de traduction automatique. La performance de ces deux types d'outils, encore très inégale entre langues dotées et langues peu dotées, est à la fois révélatrice et reproductrice des inégalités linguistiques[2].

Dans le cadre du projet, ce guide avait pour objectif, d'une part, de rendre accessible aux annotateur·ices la notion d'entité nommée et les notions liées, et d'illustrer ces notions théoriques à l'aide d'exemples en persan. D'autre part, ce guide devait transmettre aux annotateur·ices les instructions de leur mission d'annotation ainsi qu'un certain nombre de préconisations illustrées par des exemples issus du corpus persan. Il s'agit d'un document technique qui s'adresse à des personnes parlant couramment français et persan et ayant des connaissances élémentaires en linguistique.

 Nous nous inscrivons dans une perspective de promotion des langues minorées, de développement du traitement automatique des langues moins dotées en ressources et en outils numériques, et de lutte contre les inégalités linguistiques, en espérant que notre travail puisse être utile à d'autres projets de recherche dans ce domaine.

Mots-clés : entités nommées, traitement automatique des langues naturelles, langues peu dotées, farsi, persan, campagne d'annotation de corpus.

[1] EHRMANN, Maud, NOUVEL, Damien, ROSSET, Sophie (2016). Named Entity Resources - Overview and Outlook. Portorož, Slovenia. https://hal.science/hal-01359441

[2] BERNHARD, Delphine, SORIA, Claudia (2018). Traitement automatique des langues peu dotées. In : Traitement Automatique des Langues, 2018, Vol. 59, N°3. https://www.atala.org/content/traitement-automatique-des-langues-peu-dot%C3%A9es

Sommaire

I. Votre mission d'annotation

Votre mission en tant qu'annotateur·ice est de reconnaître et de catégoriser, dans un temps limité, des entités nommées (EN) dans de courts extraits en persan.

Dans cette campagne, 3 catégories d'entités nommées sont à repérer et à étiqueter :

→ les noms de lieux (= toponymes)
→ les noms de personnes (= anthroponymes)
→ les noms d'organisations (= ergonymes)

Votre corpus de travail se compose d'extraits de pages Wikipédia et de tweets. Les sources sont toutes issues de l'internet iranien et sont toutes représentatives de la variété dialectale iranienne.

II. Aperçu théorique du concept d'entité nommée

En traitement automatique des langues (TAL), on définit une entité nommée comme une unité linguistique monoréférentielle et référentiellement autonome dans un corpus donné[3].

Dans cette campagne, nous travaillons uniquement sur les trois types d'entités nommées les plus courantes, mais il en existe une très grande variété. À titre d'exemple, on pourrait également étiqueter les produits, les œuvres, les évènements, les titres, les fonctions, les valeurs numériques, les valeurs temporelles, les distances, les coordonnées de contact, les quantités, etc.[4]

II.a. Le critère de monoréférentialité

Une EN est un mot ou un groupe de mots qui fait référence à un élément à la fois existant et unique. C'est une expression qui ne possède qu'un seul référent possible. Une EN ne désigne pas un objet général mais désigne un objet en particulier : il s'agit d'une personne spécifique, d'un lieu précis, d'une

[3] EHRMANN, Maud (2008). Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL]. Université Paris Cité. Disponible sur : https://hal.science/tel-01639190

[4] GROUIN, Cyril, ROSSET, Sophie, ZWEIGENBAUM, Pierre, et al. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In : Proceedings of the 5th linguistic annotation workshop. p. 92-100. https://aclanthology.org/W11-0411.pdf

organisation en particulier, etc. Le référent d'une EN est toujours identifiable, si tant est que le contexte soit suffisamment clair pour permettre au·à la destinataire de l'identifier.

Si une EN désigne un objet existant, cet objet peut tout à fait être fictif. Toutefois, il est perçu comme existant hors du langage par l'émetteur·ice et par le·la destinataire. Par exemple, ملا نصرالدین (le mollah Nasreddine) a beau être un personnage de fiction, il est une référence claire pour les locuteur·ices iranien·nes car il existe hors du langage.

Exemple :

خانه پشت بام دارد. = La maison a un toit/Les maisons ont un toit.

En fonction du contexte, cette phrase peut avoir deux sens. En disant cela, l'émetteur·ice peut énoncer une vérité générale et exprimer que toutes les maisons possèdent un toit. Dans ce cas, « خانه » (« une maison, les maisons ») est une notion générique et n'est pas une entité nommée. Cependant, l'émetteur·ice peut également parler d'une maison en particulier, à savoir celle qu'il a sous les yeux et qui a déjà été mentionnée, éventuellement dont la localisation a déjà été précisée ou dont la description a déjà été faite préalablement, et exprimer que cette maison en particulier possède un toit. Dans ce cas, « خانه » (« la maison ») est une entité nommée.

Repérer les EN nécessite de penser en termes de référent : ce qui fait qu'une EN est une EN est son référent, c'est-à-dire l'élément existant qu'elle désigne. L'identification et la catégorisation d'une EN nécessite, en général, une certaine connaissance du monde ainsi qu'une connaissance du contexte, car le sens n'est pas tout entier contenu dans l'expression linguistique en elle-même : c'est le contexte qui donne au destinataire des indices pour comprendre à quel objet l'expression se réfère.

Une entité nommée peut être un nom propre, mais également une description définie, ce qui la rend plus complexe à identifier.

La distinction entre un nom propre et une description définie réside dans le fait qu'un nom propre nomme un objet, tandis qu'une description définie, sans le nommer, affirme qu'il existe un seul et unique objet qui satisfasse à cette description.

Exemple :

علی خامنه ای = Ali Khamenei (nom propre)

رهبر جمهوری اسلامی ایران = guide suprême de la révolution islamique d'Ian (description définie qui désigne également Ali Khamenei)

II.b. Le critère d'autonomie

Une description définie peut être plus ou moins complète. Une description définie incomplète ne fournit pas la totalité des éléments permettant l'identification du référent désigné, tandis qu'une description définie complète est autonome référentiellement : elle ne peut désigner qu'un seul objet[5].

Exemple 1 :

[5] EHRMANN, Maud (2008). Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL]. Université Paris Cité. Disponible sur : https://hal.science/tel-01639190

رئیس جمهور = le président de la République (expression définie incomplète)

رئیس جمهور ایران در سال ۱۳۹۶ = le président de la République d'Iran en 2017 (expression définie complète)

Dans les deux cas ci-dessus, l'expression désigne Hassan Rohani. Dans le 1er exemple (رئیس جمهور – « le président de la République »), l'expression est ambiguë : d'un point de vue purement linguistique, elle pourrait désigner plusieurs personnes différentes. Dans le 2e exemple (رئیس جمهور ایران در سال ۱۳۹۶ – « le président de la République d'Iran en 2017 »), l'expression est non-ambiguë et autonome référentiellement : elle ne peut désigner qu'une seule personne.

Cette autonomie référentielle est l'un des deux critères clé pour reconnaître une entité nommée. Bien que cette autonomie référentielle soit toujours relative au corpus et souvent sujette à discussion, tu dois toujours la prendre en considération dans ta mission d'annotateur·ice.

Cette autonomie référentielle peut être sujette à discussion dans la mesure où la monoréférentialité n'est réellement atteinte que contextuellement : dans le 1er exemple (رئیس جمهور – « le président de la République »), le·la destinataire doit faire intervenir autant le contexte que ses connaissances du monde pour comprendre de qui il s'agit ; quant au 2e exemple (رئیس جمهور ایران در سال ۱۳۹۶ – « le président de la République d'Iran en 2017 »), le·la destinataire doit au moins faire intervenir ses connaissances du monde.

De plus, l'autonomie référentielle est toujours relative à un corpus donné. En effet, les pronoms personnels (او، تو، من, etc.) ne sont quasiment jamais considérés comme des entités nommées de personnes, alors mêmes qu'ils désignent toujours des personnes : car, d'un point de vue purement linguistique, l'objet désigné est quasiment toujours ambigu (on ne peut pas savoir, sans contexte, qui est « je », « tu », « il »). Cependant, dans un corpus de correspondance épistolaire de Marcel Proust, on pourrait considérer que « je » est une entité nommée, dans la mesure où il n'y a qu'un seul référent et que celui-ci est non ambigu au sein de ce corpus donné.

Votre corpus de travail se compose d'extraits très courts (une phrase maximum). Vous devez donc évaluer la question de l'autonomie uniquement à l'aune de cette phrase et indépendamment des autres extraits, même si ce n'est pas toujours aisé. Vous devez étiqueter une unité linguistique à l'aune des indices présents dans la phrase si et seulement si elle vous semble désigner une entité de manière non-ambiguë.

II.c. La désambiguïsation lexicale

Certains phénomènes compliquent la reconnaissance des entités nommées. En effet, une seule expression peut être pluriréférentielle (avoir plusieurs référents selon le contexte), et, à l'inverse, une seule et même entité peut être désignée par de nombreuses expressions différentes. L'annotation exige donc un travail de désambiguïsation lexicale[6]. On peut citer :

[6] EHRMANN, Maud (2008). Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL]. Université Paris Cité. Disponible sur : https://hal.science/tel-01639190

→ la synonymie : plusieurs expressions différentes pour désigner une seule entité ;

→ l'homonymie : une seule expression pour désigner plusieurs entités distinctes au sein d'un même corpus ;

→ la métonymie : une expression associée, dans un contexte donné, à une entité différente de celle qu'elle désigne habituellement, les deux entités étant liées entre elles, par exemple la partie pour le tout ou le tout pour la partie ;

→ l'anaphore : une expression associée, dans le contexte, à une entité différente de celle qu'elle désigne habituellement, car elle répète de manière contractée une autre expression mentionnée préalablement dans le corpus, ou fait une référence implicite à cette autre expression.

Reprenons l'exemple ci-dessus :

علی خامنه ای = Ali Khamenei

رهبر جمهوری اسلامی ایران = guide suprême de la révolution islamique en Iran

رهبر = « guide »

→ Si ces trois expressions font référence à Ali Khamenei, alors elles font référence au même objet et sont donc synonymes. La première expression le fait par le biais d'un nom propre alors que la deuxième et la troisième le font par le biais d'expressions définies.

→ En fonction du contexte, l'expression رهبر جمهوری اسلامی ایران (« guide suprême de la révolution islamique d'Iran ») peut être soit une personne (si l'émetteur fait référence à Ali Khamenei), soit un titre/une fonction (si l'émetteur fait référence à la fonction qui porte ce nom, qui a été occupée par Ruhollah Khomeini puis par Ali Khamenei, et auxquels pourront succéder d'autres personnes). On dit donc qu'il s'agit d'homonymie : une même expression est associée à au moins deux entités distinctes selon le contexte d'énonciation, car elle possède des référents multiples.

→ Attention également à la métonymie. Dans la langue générale en français, lorsqu'on emploie « l'Iran » ou « Téhéran », on fait en général référence au pays et à la ville : ce sont des toponymes. Mais, dans le discours journalistique, on emploie souvent ces expressions pour faire référence, par métonymie, au gouvernement iranien, à l'Etat iranien, au régime iranien : ce sont donc des noms d'organisations (ergonymes). Il se trouve que ce type de formulations est très présent dans le discours journalistique en français, mais peu présent en persan.

→ Enfin, dans le langage courant, on désigne fréquemment par l'expression رهبر (« guide ») les objets cités plus haut (à savoir Ali Khamenei, Ruhollah Khomeini, ou encore la fonction de guide suprême). Cette expression prend généralement sens par reprise anaphorique en fonction du contexte : l'émetteur·ice et le·la destinataire expriment et comprennent implicitement qu'il s'agit du رهبر جمهوری اسلامی ایران (« guide suprême de la révolution islamique en Iran ») contracté en رهبر (« guide »). C'est d'autant plus le cas si l'expression complète a été employée préalablement dans le texte ou la situation d'énonciation. L'expression رهبر (« guide ») est alors associée à une entité différente de celle à laquelle elle est associée habituellement (celle d'un *guide* au sens général du mot) car il s'agit d'une anaphore.

Pour identifier et étiqueter les entités nommées, il faut donc penser en termes de référent et de contexte.

Attention : il ne faut pas confondre l'entité, l'entité nommée et le type d'entité nommée.

entité nommée → le mot ou groupe de mots
entité → le référent existant hors du langage désigné par ce mot ou ce groupe de mots
catégorie d'entité nommée → « noms de lieux », « noms d'organisations », etc.

III. Les trois catégories d'entités nommées à étiqueter

Dans certains cas, les contours de la notion d'entité nommée peuvent s'avérer flous, voire arbitraires. C'est notamment dû au fait que la notion est issue du terrain et non de la linguistique théorique[7]. Dans la présente campagne d'annotation, nous édictons des consignes d'annotation en nous basant sur des travaux théoriques mais également, dans certains cas, en faisant des choix régis par des impératifs pratiques. Ces choix arbitraires d'ordre technique concernent par exemple l'annotation des pronoms, l'annotation des collocations descriptives en persan, ou encore l'annotation des entités nommées imbriquées. Ces choix ont fait consensus dans notre équipe de chercheur·euses et d'annotateur·ices. Dans le cadre d'un autre projet, selon les objectifs de recherche et les moyens à disposition, les consignes d'annotation pourraient tout à fait différer.

Chacune des trois catégories est détaillée ci-dessous, avec des exemples et des contre-exemples en persan.

III.a. Les noms de lieux (ou toponymes)

Ce sont les expressions qui désignent des entités localisées géographiquement.

Vous devez annoter uniquement si le nom de lieu donne une indication géographique précise, identifie un lieu en particulier.

On peut citer les noms de continents, pays, régions, provinces, quartiers, lieux-dits, villes, divisions administratives, chaînes de montagne, mers, océans, etc.

Exemples :

دماوند = le Mont Damavand

لیختن‌اشتاین = le Liechstenstein

Les noms de lieux sont souvent des groupes de mots (=syntagmes). Ces syntagmes peuvent être des noms composés lexicalisés ou de simples périphrases descriptives. Mais dans un cas comme dans l'autre, les mots du syntagme ne peuvent être dissociés les uns des autres sans que le syntagme perde son sens.

Exemple de noms composés lexicalisés :

[7] NOUVEL, Damien, EHRMANN, Maud, ROSSET, Sophie (2015). Les entités nommées pour le traitement automatique des langues. Editions : Iste. Collection : Science Cognitive Et Management Des Connaissances. ISBN : 1784051047.

خلیج فارس = le Golfe Persique

دریای خزر = la Mer Caspienne

Exemples de périphrases descriptives :

قله توچال = le sommet du Mont Tochal

نبش خیابان موازی با بولوار انقلاب = l'angle de la rue parallèle au boulevard Enqelab

III.a.1. <u>Les éléments de lieux généraux</u>

Vous ne devez pas annoter les noms qui désignent des éléments de lieux généraux s'ils apparaissent seuls.

Exemples :

مرکز، بخش، محل، میدان، شهرستان، دریا، جزیره، خلیج، شهر، کوه، ساختمان، عمارت، پایتخت، خانه، جنگل, etc. (« forêt », « maison », « capitale », « bâtiment », « montagne », « ville », « golfe », « île », « mer », « district », « place », « département », « centre », etc.)

En effet, ce sont des éléments de lieux génériques et non spécifiques. Ils ne désignent pas un lieu en particulier.

De la même manière, vous ne devez pas annoter les points cardinaux s'ils apparaissent seuls et s'ils désignent une direction et non un lieu : شمال، جنوب، شرق، غرب (« ouest », « est », « sud », « nord »).

III.a.2. <u>Les noms et adjectifs de localisation</u>

Vous devez inclure les noms et adjectifs de localisation si et seulement s'ils font partie intégrante d'une expression qui désigne un lieu identifiable.

Les noms et adjectifs de localisation sont nombreux. On peut citer notamment les points cardinaux (شمالی، غربی، جنوبی، شرقی) mais également de nombreux adjectifs employés pour localiser : مرکزی، چپ، راست، بالا، پایین، عقب, etc. (« central », « bas », « haut », « droit », « gauche », etc.)

L'important est de toujours se demander si l'émetteur·ice désigne un lieu précis et identifiable par ses interlocuteur·ices (auquel cas vous devez annoter) ou non (auquel cas vous ne devez pas annoter).

Exemples :

آذربایجان شرقی = Azerbaïdjan oriental (nom de province)

خراسان شمالی = Khorassan septentrional (nom de province)

خاورمیانه = Moyen-Orient (nom de région)

آلمان غربی = Allemagne de l'Ouest (nom de pays)

آسیای مرکزی = Asie centrale (nom de région)

قطب شمال = Pôle Nord

Contre-exemples :

به سمت غرب = vers l'ouest

یک جای مرکزی = un endroit central

III.a.3. Les termes de localisation très courants au sein de certaines collocations

Une collocation est une cooccurrence significative, qui n'est ni un nom composé, ni une association aléatoire de mots.

Le persan à une tendance générale à la collocation descriptive.

Dans le cadre des noms de lieux, le persan fait fréquemment précéder un toponyme par un mot qui décrit le type de toponyme dont il s'agit. Cela forme des collocations, et ce premier mot s'appelle un collocat.

Lorsque ces collocations sont construites sur des mots de localisation très courants, et lorsque ces mots peuvent être dissociés les uns des autres sans que la signification ou l'identification de l'entité de lieu soit perdue, vous ne devez pas annoter le premier collocat.

Exemples :

استان مازندران = le Mazandaran

جزیره کیش = Kish

رشته کوه البرز = l'Elborz

کشور رومانی = la Roumanie

شهر ایذه = Izeh

Exception :

Ces mêmes mots de localisation très courants sont parfois inévitables, dans la mesure où ils donnent une précision importante pour distinguer une entité de son homonyme. Dans ce cas, vous n'avez pas d'autre choix que des les inclure dans l'annotation.

Exemples :

استان ایلام = la province de Ilam (et non la ville du même nom, plus connue)

شهرستان بیرجند = le district de Birjand (et non la ville du même nom, plus connue)

Notre intuition linguistique est toujours subjective. Toutefois, l'annotateur·ice doit essayer d'objectiver son intuition linguistique, en se demandant comment l'énoncé serait exprimé ou compris par n'importe quel·le locuteur·ice natif·ve moyen·ne du dialecte concerné (en l'occurrence le persan iranien).

Dans l'exemple ci-dessus, nous estimons que, par défaut, en l'absence de précisions, ایلام (« Ilam ») et بیرجند (« Birjand ») ont tendance à faire référence à des villes pour un·e locuteur·ice moyen·ne. Si le mot fait référence à la ville, vous devez donc l'annoter seul (en laissant de côté ce collocat descriptif qui le précède sans ajouter au sens). Mais si le mot fait référence à une autre division administrative du même nom, vous devez inclure ce premier collocat dans l'annotation (استان – « province » - شهرستان - « district » - etc.), car nous estimons qu'il fait partie de l'entité nommée, dans la mesure où le dissocier pourrait porter atteinte au sens et empêcher le destinataire d'identifier précisément l'entité nommée dont il s'agit.

III.a.4. Les noms de rue

Vous devez annoter l'intégralité des noms de rue comme noms de lieux.

Exemple :

بولوار انقلاب = le boulevard Enqelab

III.a.5. Les noms de lieux employés dans des noms d'organisation

Les noms d'institutions (publiques, privées, religieuses, etc.) incluent souvent un nom de lieu. Pourtant, le nom de lieu est ici mis de côté pour privilégier le nom d'organisation dont il fait partie. Vous devez donc annoter l'expression entière en tant qu'organisation, et vous ne devez pas imbriquer les entités nommées les unes dans les autres (cf. IV.a.)

Exemples 1 :

شرکت مدیریت ترافیک هوایی ایتالیایی = l'agence italienne de gestion du trafic aérien

وزارت اقتصاد و دارایی فرانسه = Ministère (français) de l'économie et des finances

بانک مرکزی ایران = Banque Centrale d'Iran

Exemple 2 :

دانشگاه کالیفرنیا، سانتا باربارا (university of california ، santa barbara)

Dans l'exemple ci-dessus, « Californie » est un nom de lieu qui fait partie intégrante du nom de l'université « Université de Californie » : le tout doit donc être annoté comme un nom d'organisation. Toutefois, « Santa Barbara » est un nom de lieu, qui est là uniquement dans le bus de localiser l'université.

Exemple 3 :

قابل توجه ترین سرودها سرود فیفا و سرود لیگ قهرمانان اروپا هستند .

Dans l'exemple ci-dessus, « européenne » est à inclure dans l'entité nommée d'organisation « Ligue des Champions (européenne) ». En effet, c'est la manière la plus idiomatique en persan de désigner la Ligue des Champions, et elle inclut la précision « européenne » qui fait ici partie intégrante de l'organisation.

III.b. Les noms de personnes (ou anthroponymes)

Ce sont les mots ou groupes de mots qui désignent des entités humaines, réelles ou fictives, contemporaines ou historiques.

Exemples :

اسکارلت جوهانسون = Scarlett Johansson

پسر ارشد پادشاه انگلستان = le fils aîné du Roi d'Angleterre

خسرو اول = Khosro 1er

III.b.1. Les pronoms personnels

Vous ne devez pas annoter les pronoms personnels.

III.b.2. Les initiales

Si une personne n'est nommée que par ses initiales (pour des raisons d'anonymisation dans les rapports de faits divers criminels par exemple), vous devez l'annoter.

III.b.3. Les identifiants Tweeter ou Instagram

Dans les tweets, si un individu est identifié par son pseudonyme (@xxxx), vous devez étiqueter cet identifiant comme personne. Il en va de même si une personne est nommée via son pseudonyme Instagram.

III.b.4. Les noms de Dieu

Vous ne devez pas annoter les noms de Dieu (خدا ،الله, etc.) et ses attributs (رحیم ،کریم, etc.)

III.b.5. Les titres et les fonctions

Les titres et les fonctions au sein des noms de personnes sont parfois à inclure, parfois non.

→ Vous devez annoter les noms précédés ou suivis d'un titre en incluant le titre, quand le titre est indissociable du nom. C'est souvent le cas de آقا (« Monsieur ») ou خانم (« Madame ») par exemple.

Exemple :

آقای آریان پور = M. Aryanpoor
خانم علی زاده = Mme Alizada

→ Les titres خان, سید, حضرت, حج (« khan », « seyyed », « hazrat », « haj ») seront à inclure dans le nom de personne s'ils témoignent d'un rang ou d'une position dans la société qui fait ou peut faire consensus.

Exemple :

سید روح‌الله خمینی = (Seyyed) Ruhollah Khomeini

حضرت ابوبکر = (Hazrat) Abubaker

→ Les noms de métiers ne sont pas à inclure dans les noms de personnes, sauf lorsqu'ils témoignent d'un grade qui fait ou peut faire consensus.

Exemple :

دکتر خدایاری = Dr. Khodayari
استاد مهدی زاده = Pr. Mehdizadeh

→ Si l'émetteur·ice présente le titre comme faisant partie intégrante du nom de la personne, vous devez aussi l'inclure dans le nom de personne.

Exemple :

در ۲ مهٔ ۲۰۱۵ ، دوشس کمبریج به بیمارستان سنت مری منتقل شد و نوزاد
دختری به نام شاهدخت شارلوت کمبریج را در ساعت ۰۸ : ۳۴ به وقت محلی به
دنیا آورد .

Dans le cas ci-dessus, le mot « شاهدخت » (« princesse ») est présenté, dans la phrase, comme faisant partie intégrante du nom de la fille nouvelle-née de la duchesse de Cambridge (« une jeune fille baptisée Princesse Charlotte Cambridge »).

→ Vous ne devez pas annoter si le titre est uniquement une épithète affectueuse, honorifique ou respectueuse, qui est uniquement le fruit que du choix de l'émetteur·ice et ne fait pas globalement consensus. C'est notamment le cas des épithètes عزیز, محترم, جناب, جان, حجی (« aziz », « mohtaram », « jenab », « jan », « hajji ») etc.

III.b.6. <u>La population d'un territoire</u>

Un groupe de personnes à la délimitation vague et peu déterminée sera considéré comme un groupe et non comme des personnes. Vous ne devez <u>pas</u> annoter « la population », « les habitants de tel endroit », « le peuple », etc.

Exemples :

شیرازیان = les chiraziens
پاریسی ها = les parisiens
مردم فرانسه = le peuple français
جمعیت مشهد = la population de Mashhad

III.c. Les noms d'organisations (ou ergonymes)

Ce sont les mots ou groupes de mots qui désignent des organisations : équipes, groupes, institutions publiques, gouvernements, ONG, entreprises, services, banques, organisations religieuses, associations, agences nationales ou internationales, partis politiques, syndicats, tribus, clans, etc.

Exemples :

حزب جمهوری خواهان = le parti républicain

مجلس فرانسه = l'Assemblée Nationale (en France)

تویوتا = Toyota

سازمان بهزیستی کشور = Sazman-e Behzisti-e Keshvar (principale organisation de charité et de travail social en Iran)

III.c.1. Les termes d'organisation très courants au sein de collocations

Comme pour les autres entités, le persan a une tendance descriptive et à faire fréquemment précéder un ergonyme par le type d'ergonyme dont il s'agit. Vous ne devez pas annoter certaines collocations construites avec des termes de types d'organisation très courants.

Exemple :

شرکت جنرال موتورز = General Motors

III.c.2. Les éléments d'organisations généraux

Comme pour les autres entités, vous ne devez pas annoter les noms d'organisations généraux s'ils apparaissent seuls.

Exemples :

پاسگاه پلیس = commissariat

شرکت عام المنفعه = société d'intérêt général

III.c.3. Attention aux noms d'organisation qui contiennent un nom de lieu

Lorsque le nom du lieu fait partie intégrante du nom d'organisation, vous devez inclure le tout dans l'étiquette « organisation » (cf. III.a.3.).

Vous ne pouvez pas imbriquer les entités nommées les unes dans les autres : vous devez choisir la plus large (cf. IV.a.).

III.c.4. Les noms de projets, d'études, de programmes

Les noms de projets, d'études, de programmes, etc. ne sont pas des organisations.

Exemples :

کمپین واکسیناسیون بیماری سل در مهاجرین = campagne de vaccination de la tuberculose chez les immigrés

پروگرام کمک های دولتی برای جامعه = programme gouvernemental d'aides financières à la communauté

III.c.5. Les sigles et acronymes en alphabet latin

Certains sigles et acronymes en alphabet latin sont employés de manière tout à fait idiomatique dans des textes écrits en persan et sont tout à fait compréhensibles pour des locuteur·ices iranien·nes moyen·nes (quelquefois même davantage que leur traduction), notamment pour faire référence à des organisations étrangères ou internationales. Le cas échéant, vous devez donc les annoter comme telles.

Exemples :

UNO = Organisation des Nations Unies
NATO = Organisation du Traité de l'Atlantique Nord

III.c.6. Le dégré de reconnaissance institutionnelle d'une organisation

Est reconnue comme organisation une entité qui a un rôle officiel et institutionnel et qui jouit d'un certain degré de reconnaissance. Un groupe de travail interne (à une institution, une entreprise, une association) ne sera pas annoté comme entité nommée, sauf s'il a un rôle officiel, institutionnel, et est largement reconnu comme tel.

Une recherche rapide sur internet peut, en règle générale, vous faire comprendre si un mot fait référence à un groupe de travail interne à une institution connu uniquement par les membres et les initié·es (à ne pas annoter) ou à un groupe de travail qui a une valeur institutionnelle reconnue aussi à l'extérieur de cette organisation (à annoter).

Exemples :

شورای امنیت = le Conseil de Sécurité (de l'ONU)

شورای اداری شهرستان کاشان = l'un des organes de décision d'un district en Iran (en l'occurrence, celui de Kashan)

کمیسیون مالی مجلس = la commission des affaires financières au sein du parlement iranien

IV. Autres instructions et préconisations pour l'annotation

IV.a. Pas de chevauchement possible

L'une des difficultés de l'exercice est que les entités nommées peuvent être emboîtées les unes dans les autres : il peut y avoir un nom de lieu au sein d'un nom d'organisation, un nom de personne au sein d'un nom de lieu, un nom d'organisation au sein d'un autre nom d'organisation, etc.

Toutefois, pour des raisons d'ordre technique, vous ne pouvez pas imbriquer une entité nommée dans une autre. Il faut sélectionner le segment le plus long, qui est aussi le thème principal de l'énoncé, et ne pas annoter les entités nommées imbriquées à l'intérieur de ce segment.

Exemple 1 :

این اتوبوس های در پلیس راه شهر سرخه جهت ثبت اطلاعات ناوگان حمل و نقل عمومی توقف دارند .

این اتوبوس های در پلیس راه شهر سرخه جهت ثبت اطلاعات ناوگان حمل و نقل عمومی توقف دارند .

Dans l'exemple ci-dessus, le thème de la phrase est « le poste de la police de circulation de la municipalité de Sorkha », à étiqueter donc intégralement comme nom de lieu. Il ne faut donc pas annoter uniquement « Sorkha », ni l'annoter à l'intérieur de l'annotation plus large.

Exemple 2 :

این شهر از پایان جنگ جهانی دوم تا زمان اتحاد آلمان شرقی و غربی در اکتبر ۱۹۹۰ ، پایتخت کشور آلمان غربی بود .

Dans l'exemple ci-dessus, on pourrait machinalement étiqueter « Allemagne de l'Ouest » comme nom d'organisation. Cependant il faut se poser la question : de quoi parle-t-on, quel est le thème principal de la phrase ? Réponse : Bonn (capitale de l'Allemagne de l'Ouest de la fin de la guerre à la Réunification). L'entité à étiqueter est donc : « capitale de l'Allemagne de l'Ouest », et non « Allemagne de l'Ouest ».

Exemple 3 :

« Son père était Reza Saqfi, le frère de Khadijeh Saqfi, qui était elle-même l'épouse de Ruhollah Khomeini. »

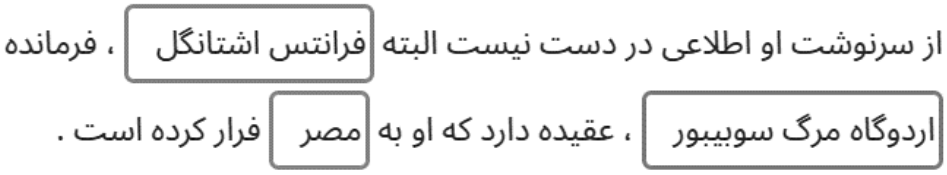
Dans l'exemple alambiqué ci-dessus, il est fait mention d'au moins trois personnages, mais la manière de désigner ces personnages emboîte les noms des personnes les uns dans les autres. Que faire ?

→ Ne pas imbriquer les entités nommées les unes dans les autres.
→ Ne pas se débarrasser du problème en étiquetant uniquement les trois noms propres (Reza Saqfi, Khadijeh Saqfi, Ruhollah Khomeini) : les noms propres ne sont pas davantage des entités nommées que les expressions définies (« le frère de Khadijeh Saqfi », « l'épouse de Ruhollah Khomeini »).
→ Identifier le thème principal de la phrase (« Reza Saqfi »).
→ Pour le reste, sélectionner les entités les plus larges (« le frère de Khadijeh Saqfi », « l'épouse de Ruhollah Khomeini »).

IV.b. Se concentrer sur la signification et la fonction du message

Beaucoup de cas posent question et font hésiter l'annotateur·ice entre plusieurs catégories. Mais en règle générale, il n'y a qu'une seule possibilité. Pour sélectionner la bonne, il faut toujours revenir à la signification et au message transmis dans le contexte d'énonciation.

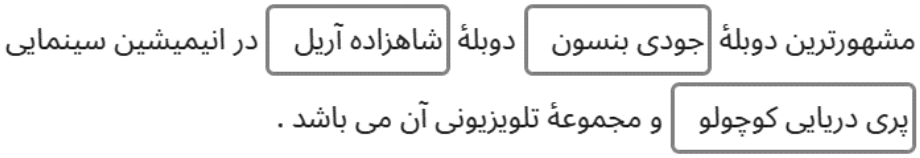
Exemple 1 :



Dans ce cas-ci, on peut se dire que « le camp de la mort de Sobibor » est un lieu. Cependant, le thème principal de la phrase est Franz Stangl, qu'on dit être le commandant de Sobibor. L'entité « Sobibor »

est donc considérée en tant qu'organisation, car on est le commandant d'une organisation et non d'un lieu.

Exemple 2 :



Dans ce cas-ci, « la princesse Ariel » est considérée comme un personnage, dont le doublage est fait par Judy Benson. Tu devras donc l'étiqueter comme un nom de personne. Toutefois, « la petite sirène » n'est pas un personnage, mais le film d'animation éponyme. S'il fallait l'étiqueter, on pourrait l'étiqueter comme œuvre, par exemple.

IV.c. Aucune règle n'est sans exception

Toutes les règles générales énoncées plus haut peuvent cacher des exceptions, car le répérage d'une EN dépend toujours du contexte. C'est pourquoi vous ne pouvez pas faire l'économie de la compréhension du contexte dans votre mission d'annotateur·ice.

Exemple :

می رویم به شمال = Je vais dans le Nord.

Nous avons dit que vous ne deviez pas annoter les points cardinaux lorsqu'ils apparaissent seuls, car ils désignent une direction. Cependant, lorsqu'un·e iranien·ne emploie le mot « شمال» (« nord »), iel peut également désigner une localité spécifique, à savoir la région au nord des montagnes Elbourz sur le bord de la mer Caspienne. Cette référence sera comprise par n'importe quel·le iranien·ne. Dans ce cas, il s'agit d'un nom de lieu identifié, et le contexte permet de déterminer que l'entité désignée est une zone géographique précise et non un point cardinal.

IV.d. Effectuer des recherches documentaires rapides

Repérer une entité nommée nécessite de comprendre approximativement la nature de l'entité en question à l'aide de sa connaissance personnelle du contexte. Toutefois, on ne peut pas connaître tous les lieux, les personnages et les organisations mentionnés dans le corpus. Les noms étrangers translittérés en alphabet arabo-persan peuvent également créer de la confusion. Même si l'objectif n'est pas de faire une recherche documentaire rigoureuse pour chaque extrait, n'hésitez pas à consulter brièvement des ressources en ligne pour vous assurer de comprendre les références dont il est question. Taper une expression dans la barre d'un moteur de recherche ou encore de Wikipédia permet souvent de clarifier ce à quoi elle fait référence. Quant à Glosbe (https://glosbe.com/fr/fa), il s'agit du dictionnaire collaboratif persan-français en ligne le plus fourni, et ses mémoires de traduction peuvent vous être utiles.

IV.e. L'étiquette « incertitude »

Durant la phase de formation, si vous ne savez pas comment annoter, notamment car vous estimez que le présent guide n'a pas été clair, vous devez annoter avec l'étiquette « incertitude ». Ceci nous permettra de discuter des cas épineux après la session, de réviser le guide, et d'harmoniser les pratiques d'annotation. Une fois la phase de formation passée, il n'y aura plus d'étiquette « incertitude » : vous devrez systématiquement trancher.

IV.f. L'absence d'entités nommées

Si le document vous semble ne contenir aucune annotation relevant de ces 3 catégories, alors n'annotez rien et validez pour passer au suivant.

IV.g. Pas d'impératif de résultat

Si, à la fin du temps imparti, vous n'avez pas eu le temps d'annoter tous les documents, cela n'a pas d'importance. Si votre annotation est incorrecte, ce n'est pas grave non plus. L'important est que vous fassiez des choix d'annotation conscients en faisant en sorte de fonder le plus possible vos choix sur les instructions énoncées dans ce guide.

V. Bibliographie

BERNHARD, Delphine, SORIA, Claudia (2018). Traitement automatique des langues peu dotées. In : Traitement Automatique des Langues, 2018, Vol. 59, N°3. https://www.atala.org/content/traitement-automatique-des-langues-peu-dot%C3%A9es

EHRMANN, Maud, NOUVEL, Damien, ROSSET, Sophie (2016). Named Entity Resources - Overview and Outlook. Portorož, Slovenia. https://hal.science/hal-01359441

EHRMANN, Maud (2008). Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL]. Université Paris Cité. Disponible sur : https://hal.science/tel-01639190

GROUIN, Cyril, ROSSET, Sophie, ZWEIGENBAUM, Pierre, et al. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In : Proceedings of the 5th linguistic annotation workshop. p. 92-100. https://aclanthology.org/W11-0411.pdf

NOUVEL, Damien, EHRMANN, Maud, ROSSET, Sophie (2015). Les entités nommées pour le traitement automatique des langues. Editions : Iste. Collection : Science Cognitive Et Management Des Connaissances. ISBN : 1784051047.