

HUFA corpus annotation guidelines

1. Annotation Criteria

The task consists of annotating 7 types of entities on clinical notes in Spanish coming from the Allergy Unit and Emergency Department of the Hospital Universitario Fundación Alcorcón. The entities belong to semantic groups of the Hospital itself, referring to treating allergies with nuts, being able to find pathologies, treatments, tests, diagnoses, pharmacology, etc. These semantic groups are as follows:

- COMORBIDITY entities describe disorders or diseases occurring in the same person. It implies an interaction between diseases that may worsen the course of both.
- MANIFESTATION refers to an indication or sign of an organic disturbance or illness. In pathology, it denotes the perceptible expression of a disease to the observer, which, once assessed, becomes a diagnostic factor.
- ALLERGY refers to entities that encompass allergies and the allergens that cause them. Distinctions should be made regarding nut allergies.
- NUT ALLERGY entity pertains to nuts and the reactions they cause in some patients, often associated with cutaneous and/or serum sensitization.
- COFACTOR entities are factors such as alcohol, narcotics, insomnia, etc. in that might contribute to the severity of a reaction independently of allergen exposure.
- PROTEIN refers to proteins that are present in different allergen sources and may be responsible for genuine sensitization and/or cross-reactivity between them.
- TREATMENT entities are measures implemented to improve, alleviate, or cure an allergy in the patient.

The objectives of annotation are to detect entities as they are mentioned in the scientific literature of the clinical notes and to standardise the entities with the corresponding medical concept. Annotation consists of marking from the beginning of the first word that makes up the entity to the end of the last word that makes up the entity. An entity may consist of one or more terms (e.g. entities are: "allergy", "anaphylaxis", "adrenaline SC").

1.1 Scope of the annotation

- When there are multiple terms that could be entities but when combined, they also form a single entity, the latter shall be annotated.

For example, the sentence "Dosis convencionales en el episodio anafiláctico" would be annotated as "Dosis convencionales en el [episodio anafiláctico] MANIFESTATION"

For example, the sentence "Se le suministra adrenalina intramuscular" would be annotated as "Se le suministra [adrenalina intramuscular] TREATMENT"

- Articles, cardinal numbers (two, three...), and punctuation marks (comma, period...) at the beginning and end of the entity will not be annotated.

For example, the sentence "Reportaron 3 episodios anafilácticos" would be annotated as "Reportaron 3 [episodios anafilácticos] MANIFESTATION"

- The start or end of the annotation cannot overlap with the start or end of another annotation. Only some entities that are completely contained within another entity are allowed (see the section on nested entity annotation).

For example, the sentence "Anafilaxia por pistacho" would be annotated as "[Anafilaxia] MANIFESTATION por [pistacho] NUT ALLERGY"

For example, the sentence "Urticária alérgica causada por alimento" would be annotated as "[Urticária alérgica] MANIFESTATION causada por [alimento] ALLERGY"

- Try to annotate as specifically as possible (the broader term) whenever it adds value to the main term.

For example, the sentence "Se le recetó corticoides orales" would be annotated as "Se le recetó [corticoides orales] TREATMENT"

For example, the sentence "Angioedema de lengua" would be annotated as "[Angioedema de lengua] MANIFESTATION"

- The complete term with the prepositional phrase is annotated if it helps understand the meaning of the text.

For example, the sentence "Se evaluaron diariamente los síntomas de infección respiratoria, diarrea aguda y efectos secundarios" would be annotated as "Se evaluaron diariamente los síntomas de [infección respiratoria] MANIFESTATION, [diarrea aguda] MANIFESTATION y [efectos secundarios] MANIFESTATION"

- The ® or TM symbols that may appear in some medications are not annotated.

For example, the sentence "Evaluar si el Quantiferon ®" would be annotated as "Evaluar si el [Quantiferon] TREATMENT ®"

- The asterisks that may appear in some proteins are not annotated.

For example, the sentence "**** a 14 (proteína 2s)" would be annotated as "**** a 14 ([proteína 2s]) PROTEIN"

1.2 Annotation of nested entities

- Entities included in broader-scope entities are annotated as long as both do not belong to the same semantic group.

For example, the sentence "LTP de avellanas" would be annotated as "[LTP de avellanas] PROTEIN, [LTP] PROTEIN de [avellanas] NUT ALLERGY."

For example, the sentence "F422 rAra h 2 Cacahuete" would be annotated as "F422 [rAra h 2 [Peanut] NUT ALLERGY] PROTEIN."

- Incorrectly pre-annotated entities should be removed.

For example, the pre-annotated sentence "El paciente presenta un cuadro de [estrés] MANIFESTATION" should have the pre-annotation removed and be annotated as "El paciente presenta un cuadro de [estrés] COFACTOR."

1.3 Annotation of discontinued entities

- Discontinuous entities are not annotated.

For example, the sentence "Manzana y pera peladas" would be annotated as "[Manzana] ALLERGY and [pera pelada] ALLERGY" The terms are separated. The entire phrase is not annotated as a single ALLERGY entity.

For example, the sentence "Eritema e inflamación en la zona de contacto" would be annotated as "[eritema] MANIFESTATION e [inflamación en la zona de contacto] MANIFESTATION." The terms are separated. The entire phrase is not annotated as a single MANIFESTATION entity.

1.4 Spelling and typographical errors

- Entities with spelling, typographical, and tokenization errors are also annotated, even if they do not follow academic or normative usage.

For example, the sentence "Náusea y vómito post operatorio" would be annotated as "Náusea y vómito [post operatorio] TREATMENT" The term "post operatorio" is annotated, even though the correct form is "posoperatorio" or "postoperatorio."

- Similarly, abbreviations with spelling errors are also annotated as long as they are not ambiguous.

For example, the sentence "Los alimentos que más cantidad de LPT suelen tener..." would be annotated as "Los alimentos que más cantidad de [LPT] PROTEIN suelen tener..." The correct form of LPT is LTP.

1.5 Abbreviations and acronyms

- Medical entity abbreviations and acronyms are annotated.

For example, the sentence "Se le suministra adrenalina SC" would be annotated as "Se le suministra [adrenalina SC] TREATMENT."

- Abbreviations and acronyms in English are also annotated, even if they are not translated or adapted to Spanish.

For example, PCR, TNF-alpha.

- Single-letter acronyms or abbreviations are not annotated.
- Abbreviations are annotated as complete words. Parts of a word are not annotated, even if they are abbreviations.

For example, the sentence "Se le suministra vitamina B12" would be annotated as "Se le suministra [vitamina B12] TREATMENT." [vitamina B] or [B] are not annotated.

- Abbreviations and acronyms of proper names are not annotated.

For example, the sentence "Consumieron dos fórmulas en cuatro ocasiones (Glucerna SR(r) Laboratorios Abbott C.A [FG] y Enterex Diabetic(r), Victus, C.A [FE])" would not have SR, C.A, FG, or FE annotated.

1.6 Negated entities

- Entities that appear in negated contexts are also annotated (the negation is not annotated).

For example, the sentence "Destacando la ausencia de edemas" would be annotated as "Destacando la ausencia de [edemas] MANIFESTATION."

For example, the sentence "No relación con cofactores: ejercicios, AINEs" would be annotated as "No relación con cofactores: [ejercicios] COFACTOR, [AINEs] COFACTOR"

1.7 Entities expressing possibility, probability, or assertion

- Terms that express modality, probability, and affirmation of a pathology or condition with respect to a circumstance, risk, or hypothetical speculation in the corpus are not annotated. Patterns that can capture these cases will be included in the algorithms.

For example, the sentence "Anafilaxia confirmada por pistachos" would be annotated as "[Anafilaxia] MANIFESTATION confirmada por [pistachos] NUT ALLERGY."

1.8 Terms in other languages

- Entities directly adapted from English, without translation into Spanish, are also annotated.

For example, the sentence "Intolerante a la lactosa tras un bypass gástrico" would be annotated as "[Intolerante a la lactosa] MANIFESTATION tras un [bypass gástrico] TREATMENT"

1.9 Anaphoric entities

- Pronouns, determiners, or adjectives that reference another entity are not annotated.

For example, the sentence "Se le suministra adrenalina como tratamiento de dicha anafilaxia" would be annotated as "Se le suministra [adrenalina] TREATMENT como tratamiento de dicha [anafilaxia] MANIFESTATION."

1.10 Failure to write down agreed terms

- "Hipersensibilidad alimentaria," "pruebas cutáneas," "sensibilización cutánea," and "agudización asmática" are not annotated.

- Recommendations that appear in clinical notes are not annotated.

For example, the sentence "mejora con esteroides tópicos e hidratación" would be annotated as "mejora con [esteroides tópicos] TREATMENT e hidratación." "Hidratación" is not annotated as a treatment because it is a recommendation.

1.11 Definition and notation of agreed terms

Immunoglobulin E (IgE) is a monomeric antibody composed of two heavy chains and two light chains. It is one of the immunoglobulins present in lower quantities in the human blood. It is involved in type I hypersensitivity reactions. IgE molecules bind to specific high-affinity receptors located on the membranes of mast cells and basophils; this binding results in the secretion of mediators responsible for the symptoms of allergic diseases.

Total Immunoglobulin E: the serum or plasma concentration of immunoglobulin E. It encompasses the sum of all IgE antibodies with their various specificities. Total IgE increases especially in allergic diseases, atopic dermatitis, hypergammaglobulinemia E syndrome, allergic bronchopulmonary aspergillosis, parasitic diseases, IgE myeloma, Wiskott-Aldrich syndrome, and graft-versus-host reaction, among others.

Specific Immunoglobulin E: the fraction of IgE antibodies that are specific to a particular allergenic source.

Therefore, based on the previous definitions, it is decided to annotate in notes as follows:

- The sentence "Inmunoglobulina E 68 UI/ml (<100)358" would be annotated as "[Inmunoglobulina E] ALLERGY, 68 UI/ml (<100)358" since it corresponds to Total Immunoglobulin E.
- The sentence "f202 Anacardo 8.20 kU/L () 5008" would be annotated as "f202 [Anacardo] NUT ALLERGY 8.20 kU/L () 5008" because it corresponds to specific IgE for cashews. The "f" is omitted as it serves as a food identifier.

2. Entity Annotation agreements

This section details some of the agreements made by the scorers following the triple annotation review.

- The sentence "Rinoconjuntivitis alérgica estacional por sensibilización a polen de gramíneas" would be annotated as "[Rinoconjuntivitis alérgica estacional] COMORBIDITY por sensibilización a [polen de gramíneas] ALLERGY." It is agreed to annotate the complete term "Rinoconjuntivitis alérgica estacional" to provide more information.
- The sentence "Alergia a Avella (angioedema labial)" would be annotated as "[Alergia a Avella] NUT ALLERGY ([angioedema labial]) MANIFESTATION" It is noted in a more specific manner to provide additional information.
- The sentence "Huevo cocido (noviembre 2021)" would be annotated as "[Huevo cocido] ALLERGY (noviembre 2021)." It is noted in a more specific manner to provide additional information.

- The sentence "Avellana, *** a 14 (proteína 2s)" would be annotated as "[Avellana] NUT ALLERGY, *** a 14 ([proteína 2s]) PROTEIN." The asterisks anonymize the term Cor. The annotators know it's Cor because the term "avellana" appears before the asterisks. It is indicated that another case could be Tri. During the anonymization process at the hospital, the term is suppressed based on the integrated rules. It is agreed to suppress the annotation of asterisks as they would introduce noise, and therefore, the decision is to omit the annotation.
- The sentence "mostaza (0mm), polen de abedul (0mm), polen de encina (0mm)" would be annotated as "[mostaza] NUT ALLERGY (0mm), [polen de abedul] ALLERGY (0mm), [polen de encina] ALLERGY (0mm)." It is agreed to annotate "mostaza" as NUT ALLERGY because it is a seed. In the list of nuts, we can also find sesame seeds. "Polen de abedul" is agreed to be annotated as ALLERGY to maintain the same structure as "polen de gramíneas" or "polen de encina."
- The sentence "Alergia a f. secos, refieren" would be annotated as "[Alergia a f. secos] NUT ALLERGY, refieren." It is annotated in a more specific manner to provide additional information.
- The sentence "Legumbres: tolera todas. Hortalizas: tolera zanahoria." would be annotated as "[Legumbres] ALLERGY: tolera todas. [Hortalizas] ALLERGY: tolera [zanahoria] ALLERGY." It is agreed to annotate following the same structure.
- The sentence "manzana y peras peladas, aguacate, ciruela." would be annotated as "[manzana] ALLERGY and [peras peladas] ALLERGY [aguacate] ALLERGY, [ciruela] ALLERGY." It is agreed to annotate all terms related to fruits that indicate whether the food includes skin or not. This is extrapolated to the shell of nuts.
- The sentence "Dermatitis seborreica dco por su Pediatra, trata con esteroides tópicos" would be annotated as "Dermatitis seborreica dco por su pediatra, trata con [esteroides tópicos] TREATMENT." It is agreed not to annotate "dermatitis seborreica" as it does not provide useful information, unlike the case of atopic dermatitis. It is agreed to annotate "esteroides tópicos" together due to the additional information it provides.
- The sentence "Inmunoglobulina E 68 UI/ml (<100)358" would be annotated as "[Inmunoglobulina E] ALLERGY, E 68 UI/ml (<100)358." It is agreed to annotate the term "Inmunoglobulina E" as an allergy.
- The sentence "Alergia a alimentos" would be annotated as "[Alergia a alimentos] ALLERGY." It is annotated in a more specific manner to provide additional information.
- The sentence "f202 Anacardo 8.20 kU/L () 5008" would be annotated as "f202 [Anacardo] NUT ALLERGY 8.20 kU/L () 5008." It is agreed to remove the "f" as it is an identifier for the food.
- The sentence "f443 rAna o 3 Anacardo" would be annotated as "f443 [rAna o 3 [Anacardo] NUT ALLERGY] PROTEIN." It is agreed to omit "f443" from the annotation as it is a laboratory identifier. "Anacardo" is annotated as a NUT ALLERGY.
- The sentence "ácaros del género Dermatophagoides pteronyssinus, epitelio de gato, hongos del género aspergillus" would be annotated as "[ácaros del género Dermatophagoides pteronyssinus] ALLERGY, [epitelio de gato] ALLERGY, [hongos del género aspergillus] ALLERGY." "Epitelio de gato" is agreed to be annotated together for more specificity.

- The sentence "22/02/2016 – pluralis 4mg/noche" would be annotated as "22/02/2016 – [pluralis] TREATMENT 4mg/noche." It is agreed to annotate "pluralis." This term is incorrectly spelled as "pluralais."
- The sentence "Jext 150 autoinyectable intramuscular" would be annotated as "[Jext] TREATMENT 150 autoinyectable intramuscular." It is agreed to annotate only "Jext" as the name of the injectable adrenaline. The term "150" refers to the amount of adrenaline and is not annotated. "Autoinyectable intramuscular" is not annotated as this information is intrinsic to the medication "Jext."
- The sentence "Durante la primavera síntomas de RC y síntomas de asma incrementados" would be annotated as "Durante la primavera síntomas de [RC] COMORBIDITY y síntomas de [asma] COMORBIDITY incrementados." Both terms are annotated as COMORBIDITY based on the context of the sentence.
- The sentence "f223 nGal d 1 Ovomucoide Huevo: 0.40 KUA/L f4 ****: 30.50 KU/L f75" would be annotated as "[[nGal d 1 Ovomucoide] [Huevo] ALLERGY] PROTEIN." It is agreed to omit "f223" from the annotation as it is a laboratory identifier.
- The sentence "LTP (rAra h 9 Cacahuete nsLTP, nPru p3 Melocoton LTP)" would be annotated as "[LTP] PROTEIN ([rAra h 9 [Cacahuete] NUT ALLERGY nsLTP] PROTEIN, [nPru p3 [Melocoton] ALLERGY LTP] PROTEIN)." It is agreed among the annotators to make this annotation.
- The sentence "No come cereales con gluten" would be annotated as "No come [cereales con gluten] ALLERGY." It is agreed to annotate the term in full to provide more information.
- The sentence "Ovoalbumina 6, ovomucoide 5, clara de huevo 8" would be annotated as "[Ovoalbumina] PROTEIN 6, [ovomucoide] PROTEIN 5, [clara de huevo] ALLERGY 8." "Ovoalbumina" and "ovomucoide" are noted as PROTEIN.
- The sentence "Igual con yogur de soja" would be annotated as "Igual con yogur de [soja] ALLERGY." Only "soja" is annotated as it is the allergen.
- The sentence "Esta primavera ha presentado picor y lagrimeo" would be annotated as "Esta primavera ha presentado [picor] COMORBIDITY y [lagrimeo] COMORBIDITY." Both terms are annotated as comorbidity based on the context of the sentence.
- The sentence "Rinoconjuntivitis crónica, asma moderado" would be annotated as "Rinoconjuntivitis crónica] COMORBIDITY, [asma moderada] COMORBIDITY." "Crónica" and "moderada" are included to provide more information to the terms.
- The sentence "positiva OVA 9mm, OVM 11mm, clara 15mm y yema 8mm" would be annotated as "positiva [OVA] PROTEIN 9mm, [OVM] PROTEIN 11mm, [clara] ALLERGY 15mm y [yema] ALLERGY 8mm." "OVA" and "OVM" are noted as PROTEIN.
- The sentence "la provocación siempre de ciego Síndrome Alérgica Oral con cacahuete" would be annotated as "la provocación siempre de ciego [Síndrome Alérgica Oral] MANIFESTATION con [cacahuete] NUT ALLERGY." "Síndrome Alérgica Oral" is included in the annotation.