instituto
de ingeniería
del conocimiento

Anotación de corpus lingüísticos: metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC)

AUTORAS (*)

Nuria Aldama
Lingüista computacional en el IIC

Marta Guerrero
Coordinadora de proyectos de PLN en el IIC

Helena Montoro
Lingüista computacional en el IIC

Doaa Samy
Lingüista computacional en el IIC

(*) Por orden alfabético de apellido

Anotación de corpus lingüísticos: metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC)

Resumen

Actualmente, los avances de Procesamiento del Lenguaje Natural (PLN) están estrechamente ligados a la Inteligencia Artificial y el entrenamiento de modelos de Machine Learning. Por ello, la demanda de corpus lingüísticos anotados que puedan servir de base para el entrenamiento de esos modelos está en auge. Con este white paper pretendemos detallar la metodología utilizada en el **Instituto de Ingeniería del Conocimiento** (IIC) para la anotación de corpus. Primero, mediante la introducción, se define qué es un corpus anotado y las tipologías que existen.

El siguiente apartado explica **la metodología MATTER**, conocida y seguida en muchos procesos de anotado. A continuación, nos acercamos a la construcción del modelo de anotación y cómo se aplica este en el proceso de anotación y en el desarrollo de las guías.

En el cuarto apartado tratamos en detalle la anotación por pares del corpus de desarrollo para extraer la información concreta que es de utilidad y que se desglosa en las guías de anotación.

Tras este apartado, explicamos cómo comprobar si las anotaciones realizadas son de calidad y fiables mediante algunas de las métricas más utilizadas para ello.

Una vez se han generado anotaciones por pares de un mismo corpus y se ha comprobado que son de calidad, queda generar el gold standard final en el proceso de armonización, explicado en el sexto apartado.

Por último, cierran este white paper unas conclusiones que sintetizan todo lo expuesto, y reflexionan y justifican la utilidad de seguir una metodología de anotación para **crear corpus anotados** de calidad en el panorama actual.

AUTORAS (*)

Nuria Aldama
Lingüista computacional en el IIC

Marta Guerrero
Coordinadora de proyectos de PLN en el IIC

Helena Montoro
Lingüista computacional en el IIC

Doaa Samy
Lingüista computacional en el IIC

(*) Por orden alfabético de apellido

instituto de ingeniería del conocimiento

Anotación de corpus lingüísticos: metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC)

instituto de ingeniería del conocimiento

01. Introducción

El lenguaje es el medio de comunicación humana por excelencia y es un rasgo distintivo de la inteligencia humana. Por eso, ha sido siempre objeto de estudio de diferentes disciplinas. Dentro del marco de la Inteligencia Artificial (IA), el **Procesamiento del Lenguaje Natural** (PLN), sea en su forma escrita u oral, presta especial atención al estudio de los fenómenos lingüísticos, sobre todo en los datos no estructurados, y a la forma en la que se realiza la comunicación humana con el fin de desarrollar soluciones prácticas que simulen las capacidades humanas cognitivas, especialmente en lo que se refiere a la comprensión y la generación del lenguaje.

Por su carácter empírico y teniendo en cuenta la base matemática y estadística de los modelos de la IA, además de la capacidad de cómputo cada vez mayor de los nuevos procesadores, las líneas actuales del PLN requieren de grandes cantidades de datos (en formato texto, audio o multimodal). Estos datos no estructurados constituyen la infraestructura lingüística necesaria para el desarrollo de aplicaciones inteligentes capaces de **tratar el lenguaje humano** en su forma oral o escrita. No obstante, estas grandes cantidades de datos lingüísticos no utilizan en su estado primitivo, sino que requieren pasar por una serie de fases de recopilación, depuración, estructuración, tratamiento, anotación, estandarización y validación para convertirse en “recursos lingüísticos” de calidad. Además, en estos últimos tiempos, con los avances de los **modelos de Deep Learning** y la especialización de estos en tareas concretas, la necesidad de utilizar conjuntos de textos anotados (corpus anotados) ha crecido, convirtiéndose en un recurso muy necesario para proceder a la automatización de tareas específicas de procesamiento del lenguaje. Los corpus anotados representan un tipo de recurso lingüístico de gran valor y utilidad para la IA y el PLN, y su desarrollo se basa en una metodología consolidada.

01.1. Corpus lingüístico anotado

Un **corpus lingüístico** se define por ser una colección de textos lingüísticos en formato electrónico para representar una lengua o variedad lingüística para el estudio o investigación a realizar (Sinclair, 2004). El corpus debe contener la parte del lenguaje que **es representativa** del aspecto concreto del mismo que se quiere analizar o automatizar. Este aspecto tiene que ir muy ligado al objetivo que se persigue al realizar la tarea de forma automática. Para automatizar, una de las opciones que tenemos es crear un corpus anotado. Entendemos por corpus

anotado al resultado final de marcar con información lingüística un conjunto de textos de forma manual.

La anotación de corpus nació para demostrar o probar una determinada teoría lingüística, y tanto era así que el corpus era fiel reflejo de la teoría que se estaba describiendo o probando (Ide y Pustejovsky, 2017). Hoy en día es habitual encontrar o crear corpus anotados cuya finalidad es servir de base para el desarrollo de modelos estadísticos o de aprendizaje automático (en inglés machine learning). En concreto, los **modelos de aprendizaje automático** aprenden a partir de ejemplos para realizar la tarea posteriormente de forma automática. Este aspecto ha condicionado tanto las etiquetas como el desarrollo del corpus de anotación, haciendo el proceso cada vez más sistemático y creando formalismos o estándares de anotación.

Pensemos, por ejemplo, que el objetivo es **crear un analizador sintáctico del español moderno**. Para ello, el corpus debe contener datos en español de los últimos años, pero para describir el corpus tendríamos que tener un mayor detalle o concreción. Por ejemplo: delimitar el periodo de tiempo (si queremos los últimos 50 años o los últimos 100), si se trata de español escrito o español hablado, así como concretar el dominio (textos administrativos, periodísticos, sanitarios, jurídicos, financieros, etc.), el registro (lenguaje formal, informal), dialectos (variantes peninsulares o dialectos hispanoamericanos), la procedencia de los datos (redes sociales, etc.), la longitud de los textos y, por último, el tamaño del corpus. Por tanto, es importante tener en cuenta, en la fase de diseño, la **tipología de corpus** para identificar qué tipo de textos se necesita y así conseguir los objetivos del estudio o del proyecto en cuestión.

01.2. Tipología de corpus

La **tipología de los corpus** se puede establecer en función de varios criterios. La figura 1 resume algunos de los criterios y los tipos de corpus más comunes. Es importante señalar que estos no son excluyentes entre sí, dado que un corpus se puede desarrollar eligiendo uno o varios de los aspectos siguientes:

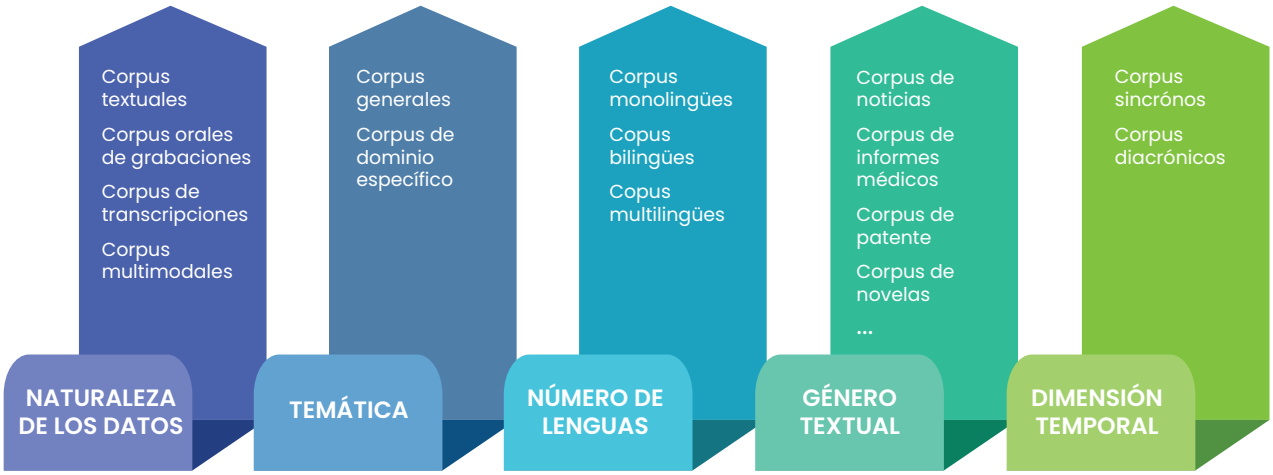

Figura 1. Tipología de corpus

Según la **naturaleza de los datos**, se puede distinguir entre corpus textuales, corpus orales, corpus transcritos (transcripciones de grabaciones de audio) y corpus multimodales. Estos últimos suelen ser corpus de vídeos que contienen texto, audio e imagen. Ejemplo de ello son los corpus que se utilizan en aplicaciones relacionadas con el procesamiento de imagen y de lenguaje, tales como traducciones al lenguaje de signos o detección de expresiones faciales asociados a ciertas expresiones lingüísticas, etc.

Teniendo en cuenta la **temática**, el corpus puede ser general, como los corpus de referencia de la Real Academia Española (RAE)[1], o de dominio específico, como los corpus de textos financieros, médicos, legales, etc. Los corpus de referencia suelen reflejar diferentes registros, variedades geográficas y una amplia distribución temática para mostrar una representación fiel de la lengua en cuestión. En cambio, los corpus de dominio específico suelen abarcar muestras representativas del lenguaje especializado y la terminología de un cierto sector.

En cuanto al número de lenguas, los corpus suelen ser monolingües, bilingües o multilingües. Estos dos últimos son imprescindibles para el desarrollo de las aplicaciones **relacionadas con la traducción automática** y asistida que suelen requerir datos en dos o más lenguas. Si se trata del mismo texto en dos o más idiomas, entonces es un corpus paralelo. Si se trata de textos parecidos del mismo dominio, pero no son iguales, entonces es un corpus comparable.

Respecto al género textual, podemos distinguir en un mismo dominio, como por ejemplo el dominio médico, entre un corpus de historia clínica y un corpus de artículos médicos, o en el dominio legal, entre un corpus de sentencias y un corpus de legislación.

El último apartado del gráfico es la **dimensión temporal**. Esta nos permite distinguir entre corpus sincrónicos que se han recopilado en el mismo marco temporal o corpus diacrónicos cuyos textos representan la evolución de una lengua, un género textual o un dominio específico a lo largo del tiempo.

Además de todo lo señalado, hay tres aspectos importantes para **diseñar el corpus**:

- **Representatividad.** En la fase de diseño es de suma importancia obtener una muestra representativa. Dado que un corpus, por muy grande que sea, nunca abarcará todo el lenguaje, sí debería ser una muestra lo más característica posible del lenguaje que se va a utilizar. Esto nos va a permitir conseguir una automatización de la tarea con mayor calidad, realizar estudios con mayor capacidad de generalización y llegar a unos resultados fiables. Por eso, si se trata de un corpus general, debería abarcar diferentes dominios y diferentes variantes geográficas, como es el caso del corpus CORDE de la RAE. Del mismo modo, un corpus oral del habla debería reflejar las variantes dialectales y debería incluir muestras representativas del habla por género y por grupos de edad, así como diferentes registros, etc. (Wynne, 2005).

- **Aspectos legales y licencias.** A la hora de la recopilación de los datos es imprescindible analizar los aspectos legales porque los hechos lingüísticos recopilados podrían contener datos personales que se deberían anonimizar. Asimismo, los textos tienen derechos de propiedad intelectual y su uso debería respetar estos derechos. Por tanto, en el proceso de recopilación es necesario asegurarse de que se recogen todos los permisos y que el proceso cumple con los derechos y las licencias pertinentes.

- **Plataformas.** Existen numerosas plataformas que ayudan en la anotación de corpus lingüísticos. Algunas de ellas están destinadas a la anotación de tareas concretas, dominios o temáticas específicas y otras ofrecen un carácter más general. Estas plataformas ofrecen muchas ventajas a la hora de realizar la anotación por varios anotadores, llevar a cabo la anotación de corpus multietiqueta, de relaciones entre las mismas o de extracción de información en textos, además, suelen impactar en una disminución de tiempos de los procesos de anotación y permiten realizar las armonizaciones de forma más eficaz, así como la evaluación del corpus final (Fort, 2016).

01.3. Anotación de corpus

La **anotación lingüística** de esos datos consiste en realizar marcas o anotaciones sobre los textos que describan, analicen o relacionen aspectos concretos. Estas marcas se hacen a nivel de palabra, sintagma, oración, fragmento o texto completo. Existen corpus anotados cuya anotación describe la morfología, sintaxis, semántica, entidades, correferencias, relaciones, etc. pero en los que la metodología general

1. Hay tres grandes corpus de referencia: CREA, CORDE y CORPES XXI. Se pueden consultar en el siguiente enlace: http://corpus.rae.es/

de anotación mantiene los principales procesos. Es frecuente encontrar, dentro de la descripción de los corpus, referencias a la **anotación por pares y métricas de acuerdo entre anotadores**. Este proceso es uno de los que **aporta calidad a la anotación**, ya que las marcas han sido acordadas por al menos dos personas. Hay otros modelos de anotación en los que hay una presencia de muchos más anotadores, donde la decisión final es llevada a cabo por una tercera figura, que puede ser un anotador experto, a veces denominado juez. Independientemente de la organización, para conseguir un corpus anotado de calidad es importante la realización del proceso de anotación por al menos dos personas. Otro de los puntos clave en el proceso de anotación es la experiencia y conocimiento técnico de los anotadores, siendo estos mencionados en muchas ocasiones en las descripciones de los corpus como garantía de buena calidad del corpus.

En este white paper nos centraremos en la metodología de la anotación de corpus en relación al Procesamiento del Lenguaje Natural (PLN) y, por lo tanto, que va a servir sobre todo para el **entrenamiento, evaluación o testeo de modelos de aprendizaje automático**. Repasaremos la metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC) para la creación de corpus anotados, empezando por la definición del modelo y pasando por las fases de preanotación, desarrollo, métricas y armonización de resultados.

02. Metodología MATTER

De entre las diferentes metodologías para la anotación de un corpus, la que seguimos en el Instituto de Ingeniería del Conocimiento (IIC) y que proponen Pustejovsky y Stubbs (2013) y Pustejovsky, Bunt y Zaenen (2017) es el ciclo de anotación denominado MATTER. Es preciso señalar que esta metodología está pensada, sobre todo, para la creación de corpus destinados a **entrenar algoritmos utilizados en tareas de PLN**, aunque existen adaptaciones para anotar corpus con otros fines[2]. Como se indica en la Figura 2, se divide en cinco pasos: modelo de anotación, anotación, entrenamiento y evaluación de un modelo de machine learning, evaluación de los resultados y revisión del modelo de anotación. Además, estas fases se pueden complementar con tres más: idea o hipótesis (previo a concretar el modelo de anotación), provisión de herramientas y formatos de anotación (previo a la anotación en sí), y distribución (una vez finalizados todos los pasos).

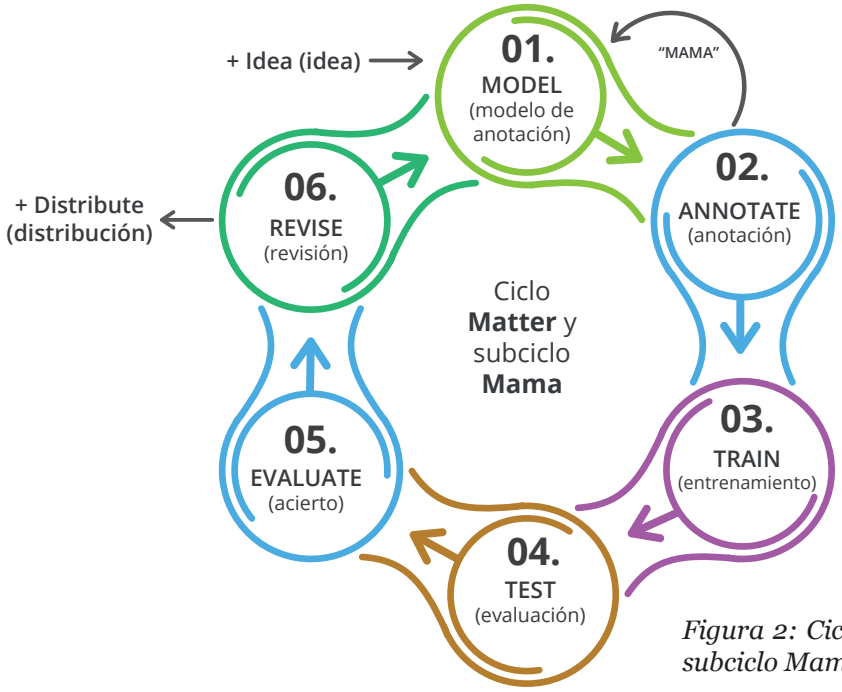

Figura 2: Ciclo Matter y subciclo Mama[3]

2. Finlayson y Erjavec (2018) proponen una adaptación del ciclo MATTER más general, pensada para describir un proceso de anotación destinada también para otros fines diferentes al entrenamiento de algoritmos en tareas de PLN.

3. Adaptada de Finlayson & Erjavec (2017) y Pustejovsky y Stubbs (2013).

Los siguientes apartados se centran en los dos primeros pasos, para explicar cómo anotar un corpus. Tanto en el primer paso ("modelo de anotación") como en el segundo ("anotación") suele producirse el subciclo MAMA (del inglés Model-Annotate-Model-Annotate) que se resume en la Figura 2. Esto implica que las tareas de creación del modelo y de anotación son cíclicas (Artstein y Poesio, 2008), de manera que las iteraciones ayudan a conseguir un anotado sólido y consistente.

03. Modelo de anotación

El **modelo de anotación** trata de encontrar un planteamiento correcto para abordar la tarea que se quiere resolver. Consiste en un marco en el que se van a definir la tarea, el fin de la misma, cómo se va a llevar a cabo, el esquema de anotación, etc., y del cual se va a partir para desarrollar los procesos que vienen después (preanotado, creación de guías de anotación y desarrollo del corpus de anotación). Algunos trabajos como los de Pustejovsky y Stubbs (2013) y Finlayson y Erjavec (2017) indican los pasos para conformar el modelo de anotación. Resumidamente, los pasos que se dan en este proceso y que seguimos en el Instituto de Ingeniería del Conocimiento (IIC) son:

1. Definir la tarea a resolver
2. Buscar bibliografía sobre la tarea y los fenómenos que se tratan en la misma
3. Recopilar los recursos que conforman el corpus
4. Obtener algunas métricas sobre el corpus que puedan ser de interés antes de anotar
5. Construir el modelo
6. Preanotar y conformar las guías de anotación

El primer punto es **definir la tarea a resolver** en una o dos frases, de manera sintética, ya que la definición del objetivo ayuda a que el modelo quede mucho más ajustado. Algún ejemplo de estas definiciones puede ser "Anotar el tipo de documento en un corpus de administración y recursos humanos para entrenar un modelo que los clasifique automáticamente" o "Etiquetar las entidades nombradas de tipo persona, organización y lugar en un corpus de noticias para entrenar un modelo que sea capaz de detectar estas entidades automáticamente". A continuación, la búsqueda de bibliografía de trabajos relacionados ayuda a conocer el estado del arte y las posibles soluciones a las dificultades encontradas en tareas similares.

La r**ecopilación del corpus** debe hacerse pensando en la tarea planteada y teniendo en cuenta la licencia de uso de esos datos. Una vez recopilado el corpus, el cuarto paso busca comprobar ciertos aspectos de los textos que pueden resultar de gran utilidad como el tamaño, volumen o distribución del corpus en base a distintas variables, como hemos visto en la introducción.

Después de estos procesos, hay que construir el modelo, preanotar un corpus y conformar las guías de anotación. Estos pasos se tratan en las siguientes secciones con más detalle.

03.1. Construcción del modelo de anotación

Llegados a este punto, ya están delimitadas una primera definición del modelo y una serie de características de la anotación. La construcción del modelo conlleva asimismo la definición del **esquema de anotación**, esto es, definir el conjunto de etiquetas que se van a utilizar, estructura, número y abreviatura de las mismas, y el formato de la anotación. Se define también si se trata de un esquema en el que las etiquetas están al mismo nivel o en el que hay **jerarquía en el etiquetario**, además de si existen o no ciertos atributos asociados a cada etiqueta.

El **esquema del modelo de anotación** se compone por una serie de términos (T), la relación entre esos términos (R) y su interpretación (I), es decir, M = <T,R,I> (Pustejovsky y Stubbs, 2013). Si se toma como ejemplo una tarea de NER que trata de "Etiquetar las entidades nombradas de tipo persona, organización y lugar en un corpus de noticias para entrenar un modelo de aprendizaje automático que sea capaz de detectar estas entidades automáticamente", se puede decir que el modelo está compuesto por:

- T = {Entidad_nombrada, Organización, Persona, Localización}
- R = {Entidad_nombrada:: = Organización | Persona | Localización}
- I = {Organización = "entidad pública o privada concreta compuesta por un conjunto de personas, regulada por una serie de normas y estructurada",

Persona = "ser humano concreto con entidad propia", Localización = "lugar físico concreto"}

Hay que tener en cuenta que la **definición y composición del modelo** puede cambiar a lo largo de todo el ciclo MATTER. Concretamente, en esta etapa se hace necesaria una reflexión que podría modificarlo: ¿qué se quiere primar en la anotación, cobertura o precisión? Es decir, hay que encontrar el punto medio adecuado entre realizar una **anotación que abarque la mayor información posible** para la tarea que se propone o que no recoja tanta información pero permita a los anotadores tener más precisión en el etiquetado. Para llegar a la combinación correcta entre cobertura y precisión, es necesario valorar el nivel de anotación y la granularidad de las etiquetas, y cómo ambas cosas impactan en la aplicación para la que se desarrolla el corpus. Es decir, **se pretende acotar el alcance del anotado** en base a los rasgos que mejor describen el fenómeno que se trata.

03.2. Preanotación y guías de anotación

En este momento, el modelo de anotación está consolidado, aunque, como en todo proceso cíclico, **puede modificarse con posterioridad**. Tal y como apuntan Pustejovsky y Stubbs (2013), los anotadores deben comprender este modelo para poder anotar el corpus con las etiquetas. Para ello, las guías de anotación tienen que recoger de manera coherente y precisa las especificaciones del etiquetado, de forma que, una vez los anotadores estén familiarizados con ellas, les ayuden en la **tarea de anotación**. Muchas veces son los propios anotadores quienes redactan estas guías, que deben ser reutilizables para posteriores procesos.

En las **guías de anotación** aparecen descritos no solo los criterios de anotación, sino algunos de los detalles que hemos visto, como la tarea (incluido los requisitos del proyecto), el nivel de la anotación, las etiquetas, el corpus con el que trabajar y el uso que se le va a dar, el formato de anotación, y el resultado final de la anotación. Tras esto, un apartado con los criterios de anotación debe recoger de qué manera etiquetar el texto.

El apartado con los criterios de anotación manual suele dividirse en cuatro tipos de reglas[4]:

1. **Reglas generales:** recogen los criterios que se aplican a todo el proceso de anotación.
2. **Reglas positivas:** recogen los criterios por los que los textos tienen que anotarse con una etiqueta específica.
3. **Reglas negativas:** recogen los criterios por los que los textos no deben anotarse con una etiqueta específica.
4. **Reglas ortográficas:** recogen los criterios ortográficos que afectan a la anotación.

Para poder añadir criterios, se establece una **primera fase de preanotado** de un subconjunto del corpus. En este primer etiquetado, varios anotadores tratan de anotar texto real por pares sin prácticamente criterios de anotación (o con muy pocos). Durante este proceso, pueden surgir dudas o problemas de anotado que los anotadores deben apuntar para incluirlos posteriormente en las guías en forma de criterios. Estos deben ser lo suficientemente generales y formales como para cubrir los **posibles casos conflictivos**.

Después, armonizamos el texto anotado, es decir, comparamos las etiquetas asignadas por los diferentes anotadores y se resuelven aquellas en las que **hay conflicto**. Por ejemplo, al armonizar la preanotación del NER que antes se comentaba, puede darse la siguiente situación:

- Anotador 1: "El [director de Colores S.L.] PER asistió ayer al acto."
- Anotador 2: "El director de [Colores] ORG S.L. asistió ayer al acto."
- Anotador 3: "El director de [Colores S.L.] ORG asistió ayer al acto."

En el supuesto anterior, cada anotador ha detectado diferentes entidades en esta frase y solo una debe ser la correcta. Los anotadores –o un juez en el caso de que lo haya– **deben decidir qué etiquetado es el correcto**, quedando recogido mediante los criterios de las guías. Si se decide que la anotación correcta es la tercera, saldrían los dos criterios siguientes de este conflicto:

1. La denominación social que acompaña a veces a las organizaciones se anota como parte de la entidad (en el apartado "Reglas positivas").

4. Esta es la agrupación de reglas que se sigue en el IIC.

2. La mención al cargo ostentado en una empresa no se anota como entidad de tipo persona (en el apartado “Reglas negativas”).

De esta manera se van recogiendo en las guías todas las dudas que surgen al preanotar, para que la anotación del corpus que después se va a realizar sea lo más coherente y precisa posible, **reduciendo los conflictos entre anotadores** y ahorrando tiempo de armonización. En ocasiones es necesario ampliar el proceso de preanotación para aclarar más dudas o completar las guías de anotación. En cualquier caso, lo que se pretende es pasar al siguiente paso del ciclo MATTER, la “A” de Annotate, con las guías de anotación cerradas y consolidadas. Sin embargo, a pesar de los esfuerzos durante la preanotación para que esto no ocurra, es común que en la anotación surjan dudas o conflictos nuevos que implican volver a actualizar las guías y el modelo, y reanotar o revisar lo ya anotado. Una de las medidas que puede ayudar a evitar conflictos es **ofrecer una formación o entrenamiento gradual a los anotadores**, para así disminuir el número de discrepancias y dudas, así como implicar un mayor número de anotadores o contar con anotadores con mayor experiencia.

En el **Instituto de Ingeniería del Conocimiento** (IIC) se sigue una metodología ágil en la que se pretende llegar a consensos y revisar el proceso durante su realización, lo cual evita dejar todas las discrepancias al final, resolviéndolas conforme avanza la anotación y permitiendo así adoptar las medidas necesarias a tiempo.

04. Anotación de corpus (de desarrollo)

La **anotación de corpus** es un proceso complejo donde un equipo de lingüistas utiliza diferentes herramientas y sistemas de anotado para extraer, subrayar o identificar información concreta de los datos. En este proceso, los anotadores, una vez “formados” y familiarizados con el modelo de anotación y las guías, asignan etiquetas a los datos en función de los criterios y guías de anotación. En este apartado, se centra la atención en las tareas principales implicadas en el **proceso de anotación**.

Durante la anotación **se aplican los criterios recogidos en las guías**. Por ejemplo, si tuviéramos un caso de clasificación automática de documentos del área de recursos humanos de una empresa, el equipo de lingüistas llevaría a cabo un proceso de anotación de corpus basado en la asignación de etiquetas (bajas, facturas, nóminas, patentes, reclamaciones...) a cada uno de los documentos que conforman el corpus (Figura 3).

Figura 3. Esquema de clasificación de documentos de recursos humanos

Para la realización del anotado del corpus se requiere que los anotadores etiqueten el mismo corpus de manera independiente, sin comunicarse entre ellos y tomando como única fuente de información las guías de anotación y los criterios definidos en ellas (Ver apartado "Preanotación y guías de anotación"). A este proceso se le llama **anotación ciega** y normalmente requiere un mínimo de dos anotadores. Esta anotación tiene tres funciones principales. Primero, garantiza la coherencia y consistencia del proceso de anotación de cara a su posible réplica en el futuro independientemente de los anotadores implicados. Segundo, comprueba que las guías de anotación contienen criterios claros que no inducen a error o ambigüedad. Tercero, al realizar la **anotación por pares**, se garantiza que no haya sesgo o un punto de vista único en la anotación de los datos.

El **proceso de anotación** puede realizarse de diferentes maneras en función de la extensión del corpus y de la complejidad de las anotaciones. Si el corpus de anotación no es muy extenso, una posibilidad es anotarlo por completo. Tras la anotación, se analiza **si ambos anotadores han asignado las mismas etiquetas** en un proceso de armonización. Si se trata de un corpus extenso, la anotación puede llevarse a cabo en lotes (batches). Esta segunda aproximación proporciona a los anotadores la posibilidad de ir resolviendo inconsistencias, dudas o casos no recogidos y modificando las guías. Además, esta estrategia trata de paliar la necesidad de reanotar fragmentos del corpus que hayan sido anotados con anterioridad a la detección del error. Este proceso de reanotación se conoce como mejora iterativa o iterative enhancement (Dickinson & Tufiş 2017).

Para llevar a cabo el proceso de anotación es importante distribuir los textos entre los anotadores. La fase de **segmentación en conjuntos de datos** de entrenamiento y evaluación, aunque no constituye una fase específica dentro del proceso de anotación de corpus de acuerdo con Pustejovsky & Stubbs (2013), es importante en cuanto al **uso de corpus anotados** en procesos de entrenamiento de modelos de aprendizaje automático. En esta fase, se divide el corpus anotado en dos conjuntos: un conjunto de entrenamiento y un conjunto de evaluación para examinar el **grado de aprendizaje del modelo**. El conjunto de entrenamiento suele constituir el 75-80% del corpus total anotado. El resto del corpus anotado (20-25%) se utiliza como corpus de evaluación. A veces, puede resultar interesante dividir los datos en conjuntos de entrenamiento y evaluación previamente a la anotación. De esta manera, se pueden realizar pruebas con el conjunto de evaluación y distintos volúmenes de datos de entrenamiento para estimar la cantidad de textos que deben ser anotados. A modo de resumen del proceso, se puede observar la Figura 4.

Figura 4. Resumen de proceso de anotación de corpus

Por último, en la tarea de evaluación se comprueba cómo de útiles y eficaces son las anotaciones para el propósito que se persigue con el corpus anotado dentro del marco donde se encuadra su uso. Algunas de las tareas relacionadas con esta fase incluyen el análisis de la distribución de las etiquetas en **el corpus anotado,** métricas de evaluación del aprendizaje (en el caso de los modelos de aprendizaje automático, como por ejemplo: precisión, cobertura, F1-score o exactitud) o cálculo de matrices de confusión que ayudan a detectar errores e inconsistencias. Si a través del proceso de evaluación se detectan errores en la anotación, falta de criterios o mejoras en el sistema de anotación, tanto el corpus completo como todo el modelo de anotación se vería sometido a **un proceso de revisión** que los situaría al inicio del ciclo MATTER de nuevo. Esto convierte a la anotación de corpus en un proceso cíclico que hace que sea costoso en cuanto a recursos humanos y tiempo.

05. Calidad y fiabilidad de las anotaciones

Como mencionan Pustejovsky y Stubbs (2013), un buen modelo de anotación que resulte en un etiquetado preciso es clave para cualquier proceso de aprendizaje automático. Por ello, tras la tarea de anotación, es necesario llevar a cabo un análisis de fiabilidad y calidad. La principal función de este análisis es la identificación de ambigüedades o inconsistencias encontradas en el desarrollo de la anotación que pueden no estar recogidas en el modelo. Además, es importante tener en cuenta que las anotaciones realizadas por varios anotadores generalmente varían y, por ello, es esencial entender los motivos por los que cada anotador ha seleccionado una etiqueta concreta y no otra en todos los casos que generan discrepancias. Como menciona Artstein (2017), que una anotación sea fiable no es condición suficiente para que sea correcta, pero sí es una condición necesaria para alcanzar un cierto grado de calidad.[5] El diagrama de la figura 5 explica el flujo de trabajo hasta obtener el grado de fiabilidad deseado en las anotaciones antes de proceder con la anotación del resto del corpus.[6]

Una de las maneras de evaluar la fiabilidad de las anotaciones es calcular **el acuerdo entre anotadores**. Existen varias formas de calcular el acuerdo entre anotadores. Por una parte, se encuentran las métricas basadas en el cálculo de los porcentajes de acuerdo observado. Entre estas métricas está el acuerdo bruto entre anotadores, una de las métricas más usadas, que se obtiene calculando el número de elementos (textos, palabras, elementos a anotar en el corpus) que reciben etiquetas idénticas por parte de los dos anotadores (Fort, 2016).

$$\frac{\text{nº de casos con etiquetas idénticas}}{\text{nº de casos totales a etiquetar}} \times 100$$

Figura 6. Cálculo de acuerdo bruto entre anotadores

Las principales ventajas que presenta esta métrica (Figura 6) son la posibilidad de llevar a cabo el cálculo sobre cualquier tipo de anotación y la alta frecuencia con la que se utiliza en los procesos de calidad de anotación de **corpus en la literatura**. Esto último facilita un marco comparativo común para diferentes proyectos en los que se lleva a cabo anotación de corpus.

No obstante, esta métrica de calidad tiene ciertas limitaciones, ya que no toma en consideración la distribución entre las **categorías de anotación**. Por un lado, favorece un mayor grado de acuerdo en categorías más comunes y, por otro, difumina las discrepancias que se dan en casos pertenecientes a

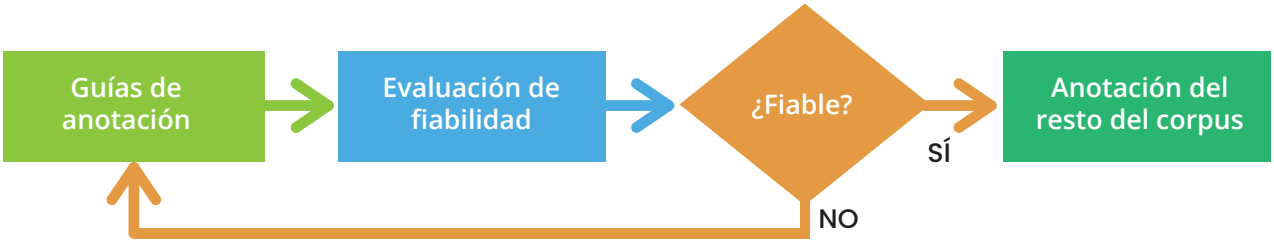

Figura 5. Flujo de tareas en función de la evaluación de las anotaciones

5. Según explican Fort (2016) y Artstein (2017) no es posible evaluar la corrección de las anotaciones llevadas a cabo por los anotadores puesto que los criterios de anotación no siguen un estándar, sino que son elaborados y ajustados a las necesidades de la tarea concreta que se persigue con la anotación de corpus.

6. En el caso representado en el diagrama, se toma una parte del corpus para evaluar la calidad y fiabilidad de las anotaciones, y cuando se obtiene un nivel de fiabilidad suficiente, el proceso de anotación puede llevarse a cabo incluso por un único anotador hasta completar la anotación del corpus.

las categorías menos representadas. Además, el acuerdo bruto entre anotadores no tiene en cuenta aquellos casos coincidentes debidos al azar.

Por otra parte, se encuentran las **métricas de calidad correctivas**,[7] que añaden a las métricas de acuerdos observados un coeficiente para indicar la probabilidad de que algunos de estos acuerdos sean debidos al azar (Chance-Corrected Coefficients). Dentro de las métricas correctivas de acuerdos esperados por azar se pueden distinguir dos subclases: los coeficientes basados en la probabilidad de "acuerdos" o coincidencias y los **coeficientes basados en la probabilidad** de "desacuerdos" o discrepancias. Los coeficientes basados en la probabilidad de "acuerdos" más comunes son π de Scott y κ de Fleiss (1971). Todos ellos parten de la misma base, que consiste en calcular la diferencia entre los acuerdos esperados por azar y los **acuerdos reales observados**. Los coeficientes más comunes que tienen en cuenta los "desacuerdos" observados son α de Krippendorff y Kappa ponderado κw de Cohen. Principalmente se basan en penalizar en mayor o menor medida los casos de desacuerdo en función de las etiquetas elegidas.

La principal desventaja que presentan tanto el porcentaje de acuerdo observado como las métricas correctivas según Fort (2016) es que son únicamente apropiados para tareas en las que se conoce el número de segmentos a anotar (p.e. procesos de tokenización, etiquetado morfológico o sintáctico en el que cada token lleva asociada una etiqueta). Sin embargo, estos coeficientes no son apropiados para tareas en las que, a priori, no se puede saber el número de elementos que será anotado (p.e. detección de entidades, clasificación multietiqueta). Para estos casos, es necesario contar con métricas como el índice de Jaccard o coeficiente de similaridad de Jaccard (Jaccard, 1912), que es una estadística utilizada para medir la similitud (y diversidad) de conjuntos de muestras. Aplicado a la anotación de corpus, el índice Jaccard calcula el número de etiquetas coincidentes proporcionadas por los anotadores para cada elemento de la anotación, independientemente del número total de etiquetas aplicadas al corpus en su versión final.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Figura 7. Cálculo de índice de Jaccard

Además de utilizar métricas para evaluar la calidad de las anotaciones, es importante tener en cuenta que hay diversos factores que **incrementan la complejidad** de evaluación del acuerdo entre anotadores:

- **Diversidad de los datos.** Un factor a tener en cuenta es la homogeneidad de los datos a anotar. Si dentro del corpus de anotación se encuentran datos de diferente naturaleza y complejidad, las métricas de evaluación de calidad han de medirse para cada subgrupo que conforme el corpus total.
- **Similitud de las etiquetas.** El segundo factor a tener en cuenta es la similitud entre las etiquetas utilizadas para anotar. Las etiquetas que pretenden recoger fenómenos muy cercanos entre sí son fuente de discrepancia puesto que, cuanto menos formal y más sutil es el grado de detalle que dirige la elección entre una etiqueta u otra, mayor será el grado de desacuerdo o de acuerdo por azar ligado a esas etiquetas.
- **Dificultad para caracterizar las etiquetas.** Las etiquetas utilizadas en un proceso de anotación han de situarse en un plano equivalente de dificultad y complejidad tanto a la hora de definirlas como a la hora de aplicarlas. La variabilidad en el grado de complejidad puede llevar a una aplicación sesgada de las etiquetas.
- **Diferencias entre los anotadores.** Las diferencias de conocimiento del mundo y conocimiento específico del dominio concreto (dominio legal o biomédico). Además, es importante considerar las diferencias de formación a la hora de enfrentarse a una tarea de anotación, ya que si los anotadores no se encuentran igualmente preparados, las anotaciones se verán afectadas y, por tanto, la calidad de las mismas.
- **Consistencia en las anotaciones de cada anotador.** A los retos de sesgo y discrepancias comunes en el proceso completo con todos los anotadores, se les puede añadir los sesgos y discrepancias individuales a nivel de las anotaciones realizadas por un mismo anotador, ya que con el tiempo un anotador puede cambiar su criterio.

7. Ver Fort (2016) y las referencias allí contenidas.

Pese a las buenas prácticas que establecen mínimos necesarios de acuerdo para la anotación, siempre se ha puesto en valor un grado de flexibilidad a la hora de interpretar las métricas y establecer criterios y parámetros. Esta flexibilidad se fundamenta en la **importancia de valorar el caso** concreto a tratar y la tarea de anotación en cuestión, porque, dada la complejidad, y la ambigüedad del fenómeno lingüístico en algunas tareas, en ocasiones es imposible llegar a un alto grado de acuerdo antes de la armonización.

06. Armonización de las anotaciones

Partiendo de las métricas y las dificultades identificadas en el proceso anterior de anotación, siempre se requiere de una fase de armonización. Por armonización, nos referimos al **proceso de resolver las discrepancias** y evitar los posibles sesgos en la anotación para conseguir tener un corpus anotado unificado, que se suele denominar corpus gold standard. Según Pustejovsky y Stubbs (2012), un corpus gold standard es una versión final armonizada y validada de un corpus etiquetado. Esta versión final se ha etiquetado según los criterios de las últimas guías de anotación y, finalmente, se ha revisado y armonizado tomando las decisiones pertinentes en caso de discrepancias.

Para afrontar la armonización, existen dos aproximaciones. Una **aproximación cuantitativa** que se basa principalmente en las métricas mencionadas en el apartado anterior para garantizar la consistencia y fiabilidad (Carletta, 1996), y otra **aproximación cualitativa** que pretende centrarse en la naturaleza de cada tarea independientemente de los resultados de las métricas (Amidei et al., 2018; Klenner et al., 2020).

En la primera aproximación, nos encontramos con las conclusiones de Carletta (1996) y todos los seguidores de esta línea que consideran que debe haber un mínimo de acuerdo de fiabilidad y consistencia dependiendo de la tarea en cuestión. Mientras que unos valores inferiores a estos mínimos establecidos solo servirían para sacar “conclusiones provisionales”.

En la segunda, han surgido varias voces en la comunidad científica reclamando una nueva lectura de estas métricas basadas en criterios cuantitativos (Amidei et al., 2018; Klenner et al., 2020). Klenner et al (2020) reclaman que centrarse en llegar al máximo nivel de acuerdo puede dañar en algunos casos el proceso de creación de recursos, ya que puede afectar a la representatividad del lenguaje natural donde la variabilidad es un rasgo fundamental. Estas opiniones muestran sus **reservas acerca de la simplificación** que supone reducir algunos fenómenos lingüísticos a una mera cifra de acuerdo. Estos fenómenos, por la propia naturaleza del lenguaje, pueden implican cierta complejidad o riqueza por la variabilidad de su uso y sus interpretaciones. Incluso Amidei et al (2018) señalan que en ocasiones la generación de recursos con la intención de **minimizar el desacuerdo** entre los anotadores puede resultar en que un corpus se distancie de las propias características del lenguaje natural, convirtiéndolo en una muestra no representativa o de un lenguaje no-natural.

06.1. Mecanismos de armonización

Una vez definido el marco de la armonización, es importante decidir el proceso que se va a llevar a cabo. Para ello, se plantean las siguientes cuestiones: ¿cuáles son los escenarios que se pueden adoptar para realizar la armonización?, ¿cuáles son las ventajas y desventajas de cada escenario?

Ante la falta de los mínimos necesarios de acuerdo entre los anotadores, nos encontramos ante dos posibles escenarios: que sea necesaria una **revisión completa** del proceso o realizar una **armonización de los textos.**

Escenario 1: Revisión completa. Se trata del caso en el que no se ha alcanzado el grado mínimo de acuerdo entre los anotadores y, por tanto, no es posible avanzar con el proceso. En este caso habría que identificar cuál es el motivo, centrándose en las siguientes posibles razones:

- La inviabilidad del modelo en sí mismo y el esquema de anotación adoptado para materializarla.
- La ambigüedad de las guías de anotación y criterios acordados.
- La falta de entrenamiento o carencias en el perfil de los anotadores, dado que algunas tareas requieren un anotador experto y otras tareas pueden incluir un perfil más genérico, como aquellas que acuden a crowdsourcing (por ejemplo “amazon mechanical turks”).

Escenario 2: Armonización. Si el acuerdo alcanzado está en un rango medio que no requiere una revisión completa del proceso, se puede pasar a un proceso de armonización. Es posible que desde el principio se haya decidido crear un gold standard con todos los textos anotados proporcionados al inicio, pero también es habitual que se acuerde tener un **consenso entre los anotadores** de un porcentaje de los textos iniciales, es decir, conseguir un gold standard con un número mínimo necesario de textos. En el Instituto de Ingeniería del Conocimiento (IIC) habitualmente se armonizan todos los textos que ofrecen discrepancia, y no solo los necesarios hasta llegar al mínimo establecido.

Para realizar esta armonización existen diferentes posibilidades:

1. Analizar los casos de desacuerdo entre los anotadores.
2. Votar por mayoría entre los anotadores.
3. Asignar los casos de desacuerdo a un experto.
4. Asignar los casos de desacuerdo a nuevos anotadores.

Todos los mecanismos señalados son válidos y la decisión de optar por uno o por otro depende de varios factores que pueden ser extrínsecos al proceso de anotación (como la disponibilidad de tener anotadores expertos o de contar con recursos adicionales). También pueden ser factores intrínsecos propios de la tarea o la teoría lingüística en cuestión (como la experiencia y el dominio de la teoría). Además, cada mecanismo tiene sus ventajas y desventajas.

En primer lugar, el procedimiento de **analizar los casos** de desacuerdo entre los anotadores que han anotado el corpus es un mecanismo eficiente y no requiere de anotadores adicionales ni de expertos. Además, es utilizado en el caso de **anotación por pares**, llegando a un consenso más ágil en cuanto a los criterios establecidos y la aplicación de las guías de anotación.

En segundo lugar, el procedimiento de votar por mayoría es el más común por su eficiencia y poco coste. Sin embargo, con este mecanismo se puede correr el riesgo de que el recurso generado (gold standard) no sea un reflejo representativo de la realidad porque no se basa en **un análisis detallado** de los motivos detrás del desacuerdo ni pretende especificar nuevos criterios para estos casos. Por el contrario, se decanta por un criterio más cuantitativo que puede ser fruto de cierto sesgo propio del proceso de anotación y no del fenómeno en cuestión, convirtiendo el **corpus** en un recurso "artificial" y "mecanizado".

En tercer lugar, asignar los textos de desacuerdo a un experto es también una práctica común que en algunos casos es necesaria, fructífera y fiable, pero siempre en la medida adecuada, ya que una excesiva dependencia de un experto conlleva otro sesgo individual e impide la transferencia y la replicabilidad del esquema de anotación. Por eso, se recomienda recurrir a un experto para resolver casos puntuales.

Por último, ampliar el número de anotadores es una solución respaldada por la comunidad científica, especialmente desde el punto de vista de las métricas porque cuando el número de anotadores es mayor, el sesgo disminuye y permite una mejor normalización de las métricas. Sin embargo, **recurrir a nuevos anotadores** requiere la disponibilidad de recursos, dado que este mecanismo implica mayor coste, más tiempo y más anotadores disponibles con perfiles adecuados. Además, la necesidad de nuevos anotadores requiere otra fase de entrenamiento del proceso de anotación, puesta en común sobre la tarea y adquisición del conocimiento contenido en las guías de anotación. Por lo que, de nuevo, puede conllevar más tiempo, más esfuerzo y más coste.

07. Conclusiones

En este white paper se detalla la **metodología para realizar la anotación de corpus lingüísticos** utilizada en el Instituto de Ingeniería del Conocimiento (IIC). Se propone una serie de fases o procesos para realizarlo siguiendo los objetivos de calidad del corpus anotado, agilidad en los procesos y revisión de los criterios acordados. A modo de resumen, en la fase inicial se describe la definición del modelo y del objeto de la tarea a anotar. Una vez empezado el proceso de anotación, se proponen las siguientes fases: fase de descripción del modelo lingüístico y preanotación del corpus, donde se va a definir el modelo lingüístico inicial y a anotar un subconjunto de textos por varios anotadores. Una vez conseguidos ciertos consensos en el proceso y después de haber realizado una armonización del corpus de preanotación, se pasa a la anotación del corpus de desarrollo. La anotación de este corpus culmina en la revisión de la calidad y fiabilidad de la anotación a través de métricas específicas. Como fase final tenemos el proceso de armonización de la anotación de corpus, que da lugar al gold standard.

Tomando como referencia el contexto actual del **Procesamiento del Lenguaje Natural** (PLN), donde hay una verdadera explosión de los modelos del lenguaje tipo BERT, RoBERTa, GPT-3, etc., el uso de corpus anotados es clave para automatizar tareas específicas como, por ejemplo: análisis del sentimiento, detección de entidades, anonimización, clasificación temática, entre otras. Para realizar estas adaptaciones a tareas concretas, conseguir recursos lingüísticos anotados de calidad es esencial. Así, resulta realmente necesario, para conseguir calidad en la automatización de tareas, contar con una metodología de **anotación de corpus lingüísticos** que aplique de forma sólida y coherente cada fase y que, a la vez, sea lo suficientemente flexible para poder abordar distintas tareas y responder a diferentes objetivos.

Por lo mencionado anteriormente, la definición de procesos y la estandarización del trabajo de anotación de corpus adquieren cada vez un **papel más relevante**. En estos últimos tiempos encontramos en los distintos catálogos de recursos online una mayor presencia de las guías de anotación que acompañan al corpus, dotando a este de una mayor transparencia y calidad. Por otro lado, se observa además un auge de competiciones y procesos de evaluación de **sistemas de machine learning** de distintos dominios, donde se valora mucho la existencia de corpus anotados manualmente de calidad.

Un factor fundamental para la anotación de corpus es el equipo de personas que hay detrás del proceso. En muchas ocasiones se menciona la **experiencia de los anotadores** como sello de calidad. Esto es así porque conseguir tener claras las fases de la anotación así como experiencia en las mismas, además del **conocimiento lingüístico** necesario para llevar a cabo las tareas, hace que el resultado final tenga mayor calidad y, por tanto, nos dirija a tener una base mejor sobre la que realizar las siguientes etapas de entrenamiento del modelo.

Por último, una metodología enfocada a la **realización de la anotación en espacios de tiempo cortos**, donde se fomente la agilidad así como la revisión y el consenso de los criterios, aporta seguridad y calidad en el proceso completo de anotación, reduce tiempo y costes, y nos dirige hacia una mayor confianza en el resultado final.

08. Referencias

Amidei, J. Piwek, P. and Willis, A. (2018). Rethinking the agreement in human evaluation tasks. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, Volume 34, (555-596).

Artstein, R. (2017) Inter-annotator Agreement. In Ide, N. and J. Pustejovsky (Eds.) Handbook of Linguistic Annotation. Springer, Dordrecht (297-314).

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2):249–254.

Carletta, J. Amy I., Stephen I., Jacqueline C., Kowtko, G., Doherty-Sneddon and Anne H. Anderson. (1997). The reliability of a dialogue structure coding scheme. Computational Linguistics, 23(1):13–32.

Dickinson, M. & Tufiş, D. (2017) Iterative Enhancement. In Ide, N. and J. Pustejovsky (Eds.) Handbook of Linguistic Annotation. Springer, Dordrecht (257-271).

Finlayson, M. A. & Erjavec, T. (2017) Overview of Annotation Creation: Processes and Tools. In Ide, N. and J. Pustejovsky (Eds.) Handbook of Linguistic Annotation. Springer, (167-192).

Fleiss, J.L. (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76 (5).

Fort K. (2016) Collaborative annotation for reliable natural language processing. In: Technical and Sociological Aspects, 1st edn. Wiley- IEEE Press.

Hovy, E. H. and Lavid, J. M. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. International Journal of Translation 22: 1, Jan-Dec. 2010: 13—36.

Ide, Nancy, Pustejovsky, James (Eds.) (2017) Handbook of Linguistic Annotation. Springer Netherlands.

Ide, N. et al. (2017). Designing Annotations Schemes: From Model to Representation. In Ide, N. and J. Pustejovsky (Eds.), Handbook of Linguistic Annotation. Springer, Dordrecht (73-72).

Jaccard, P. (1912) The distribution of the flora in the Alpine Zone 1. New Phytologist. 11 (2).

Klenner, M., Göhring, A., Amsler, M. (2020). Harmonization sometimes harms. En: S. Ebling, D. Tuggener, M. Hürlimann, M. Cieliebak, M. Volk (Eds.), Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020.

Krippendorff, K. (1978). Reliability of Binary Attribute Sata. Biometrics, 34(1):142–144. Letter to the editor, with a reply by Joseph L. Fleiss.

Krippendorff, K (1980). Content Analysis: An Introduction to Its Methodology, chapter 12. Sage, Beverly Hills, CA.

McEnery,T & Hardie,A.(2011) Corpus Linguistics: Methods, Theory and Practice.Cambridge Textbooks in Linguistics

Pustejovsky, J. and Stubbs, A. (2012). Natural Language Annotation for Machine Learning:A Guide to corpus-building for applications. O'Reilly Media, Inc.

Pustejovsky, N., Bunt, H. & Zaenen, A. (2017). Designing Annotations Schemes: From Theory to Model. In Ide, N. and J. Pustejovsky (Eds.), Handbook of Linguistic Annotation. Springer (21-111).

Sinclair, John (2004) Trust The Text. Language, Corpus and Discourse. Routledge (Taylor and Francis).

Wynne, M (Ed.) (2005). Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books.

Síguenos en:

C/ Francisco Tomás y Valiente, nº 11
EPS, edificio B, 5ª planta
UAM Cantoblanco
28049 Madrid, España.

Tel.: (+34) 91 497 2323
Fax: (+34) 91 497 2334
iic@iic.uam.es
www.iic.uam.es