**Annotation guidelines (BenchIEFL)**

This document contains annotation guidelines for the open information extraction task and has been used in the annotation process for the BenchIEFL reference. The various principles dictate which facts should and should not be annotated. Examples in green represent facts that should be annotated, while examples in red should not. In this document we refer to annotated facts by using the words tuple, fact, and triplet interchangeably. The information is presented in the following format: The different formulations of the same cluster (of the same fact), are in one paragraph and a line break separates them.

| *Phrase* | Cluster 1 Formulation 1  Cluster 1 Formulation 2  Cluster 2 Formulation 1  etc. |
| --- | --- |

**Number of arguments**

All tuples must contain between 1 and 2 arguments. Extractions with more than two arguments can be split into more compact extractions.

| *Kyle left for school on Monday* | (Kyle, left for, school)  (Kyle, left for school on, Monday)  (Kyle, left on, Monday)  (Kyle, left for, school, on Monday) |
| --- | --- |

**Relevant information**

Annotated tuples must contain relevant information that is expressed in the sentence. Tuples must be informative and relevant. They must not contain generality or empty words that convey no information. In the example, the fact that *he has written* is not relevant since it is a generality, most people *have written* and it is not the information presented in the sentence.

| *He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.* | (He, has written, opinion pieces)  (He, has, written) |
| --- | --- |
|

**Minimality**

Annotated tuples must contain minimal information, which cannot be subdivided into smaller pieces of information. No argument should contain information about two different entities if this is true for both, and no tuple should contain more than one piece of information about an entity if these can be divided. In some specific cases, such as the third example, certain relations like *combines* do not lend themselves well to minimality, so it is necessary to create versions of the annotation, one where the arguments are separated into two different clusters (elements of a city and elements of a prefecture) and one where the two arguments are grouped together. This second version is added to the two previous clusters. In this way, two systems that separate or group arguments will only be distinguished by a slightly lower accuracy for the system that separates the extractions.

| *Bob is a canadian musician* | (Bob, is, **canadian**)  (Bob, is, **a musician**)  (Bob, is, **a canadian musician**) |
| --- | --- |
| *The group was created in 2020 by three people* | (The group, was, created)  (The group, was created in, 2020)  (The group, was created by, three people)  (The group, was, created in 2020 by three people) |
| *Tokyo is often referred to as a city, but is officially known and governed as a "metropolitan prefecture", which differs from and combines elements of a city and a prefecture, a characteristic unique to Tokyo.* | (Tokyo, combines, elements of a prefecture)  (Tokyo, combines, elements of a city and a prefecture)  (Tokyo, combines, elements of a city)  (Tokyo, combines, elements of a city and a prefecture) |
|

It is sometimes necessary to separate information, if and only if it is also true when separated. In the second example, the dog is neither *black* nor *brown*, but *black* AND *brown*, whereas in the first example, *He* has *Cornish ancestors* AND *He* has *Welsh ancestors*.

| *He has Cornish as well as Welsh ancestry.* | (He, has, Cornish ancestry)  (He, has, Welsh ancestry) |
| --- | --- |
| *The dog is black and brown.* | (The dog, is, black **and** brown) |

**Exhaustivity**

All the minimal tuples presented in the sentence must be included in the annotations. Some arguments or relations may be affected by modifications but remain true without them, so it's necessary to list all possible formulations that respect the other principles. In the example, it is true that *he* *wrote* *opinion pieces*, *newspaper opinion pieces* and *magazine opinion pieces*, so all these facts must be listed in three separate clusters.

| *He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.* | (He, has written, several newspaper opinion pieces)  (He, has written, several magazine opinion pieces)  (He, has written, several opinion pieces) |
| --- | --- |

**Relation completeness**

Relations are the vehicles of information; arguments must not contain information that changes their meaning. Relations can be complicated but necessary, while they can sometimes be simplified. They must be simplified as much as possible to respect the principle of minimality, without losing their original meaning, expressed in the sentence. In the example, the second argument of the erroneous annotation contains the word *over*, which modifies the meaning of the relation *is*, whereas in the positive example, the second argument, *13 Millions*, is only the object of the relation.

| *Tokyo’s population is over 13 Millions* | (Tokyo’s population, **is over**, 13 Millions)  (Tokyo’s population, **is**, over 13 Millions) |
| --- | --- |

Sometimes, relationships can be complicated but necessary, while sometimes they can be simplified, keeping the additional part optional only if it's made necessary by the lack of other tuples explaining that additional part, as in the first example where the part *from Hungary* in the second cluster is optional because the place of origin of their escape is present in the first cluster. In the second example, *in Paris* is not optional because without this information, the relationship no longer holds.

| *His parents are Ashkenazi Jews who had to flee from Hungary during World War II.* | (His parents, had to flee from, Hungary)  (His parents, had to flee from Hungary during, World War II)  (His parents, had to flee during, World War II) |
| --- | --- |
| *Chilly Gonzales is a Grammy-winning Canadian musician who resided in Paris, France for several years, and now lives in Cologne, Germany.* | (Chilly Gonzales, resided in Paris for, [several] years) |

**Correference**

No coreference resolution is performed outside sentences. Even if a given sentence comes from a document that allows us to resolve a coreference, as OIE is intended to be a task performed on isolated sentences, we only resolve the coreferences of entities included in sentences taken in isolation. Tuples using pronouns for which we can't identify the substitution element may seem meaningless, but coreference resolution must take place outside OIE, being a task in itself. In the example, we don't do coreference resolution for the pronoun *He*, as no information about it is available in the sentence. However, we include a formulation replacing *them* with *tax reductions* in the annotation.

| *He did not go as far as he could have in tax reductions ; indeed he combined them with increases in indirect taxes .* | (He, combined **them** with, increases in indirect taxes)  (He, combined **tax reductions** with, increases in indirect taxes) |
| --- | --- |

**Inference**

Inference is necessary: facts directly implied by the sentence, even if not expressed verbatim, are relevant pieces of information. A nuance is necessary here with regard to potential implicit facts. These are not necessarily implied by the sentence and should therefore be omitted. In the first example, it is necessarily implied by the sentence that *Paul Johanson* is *Monsanto's Director of Science*. On the other hand, the fact that *their spray* is g*entle on the female organ* is not necessarily true, what is true that this information is said by *Paul Johanson*.

| *However , Paul Johanson , Monsanto ’s director of plant sciences , said the company ’s chemical spray overcomes these problems and is gentle on the female organ .* | (Paul Johanson, is, Monsanto's director of science)  (Paul Johanson, says, the company’s chemical spray is gentle on the female organ)  (the company’s chemical spray, is gentle on, the female organ) |
| --- | --- |

It is then necessary to distinguish between light and heavy inference. We define light inference as a form of inference that does not require logical reflection with respect to the sentence to deduce the fact, which is simply true as long as the sentence is also true. In the first and fourth examples, the relation is implicit, but the annotated fact is obviously true. Heavy inference, on the other hand, requires some reflection or combination of logical operations to imply the fact. A case of heavy inference is that which requires external knowledge, as in the second example, where knowledge of human culture and the principle of heredity is necessary to make the inference. Another example of heavy inference is generalization, as in the third example, where a stronger fact is implied, a generalization of what is expressed in the sentence using a single example. A final example of heavy inference is that of lower or upper limits. As in the fifth example, we don't want to generalize lower or upper bounds to entities that are not directly expressed in the sentence. We therefore include in the reference facts that can be inferred using light inference, but not those resulting from heavy inference.

| *Jason Charles Beck, a Jewish Canadian musician, was born in 1972.* | (Jason Charles Beck, is, Jewish) |
| --- | --- |
| *Gonzales is the son of Ashkenazi Jews who were forced to flee from Hungary during World War II.* | (Gonzales, is, Jewish) |
| *Gonzales is a McGill-trained virtuoso pianist.* | (McGill, trains, pianists) |
| *Nomura is amongst the more bullish.* | (Nomura, is, bullish) |
| *The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.* | (Arthur Dempster, wrote, a classic paper) |
| *The prefecture is part of the world's most populous metropolitan area with upwards of 37.8 million people and the world's largest urban agglomeration economy.* | (the world, has upwards of, 37.8 million people)  (the prefecture, has upwards of, 37.8 million people) |

Another detail is that facts must be included even if they can be inferred by a combination of logical operations on other facts. Indeed, it is generally necessary to know the language and workings of the world to make this kind of inference, which cannot be expected of an OIE system. In the first example, the third cluster can be implied by a combination of the first two, but we include it anyway, as this combination operation is not expected of an OIE system. In the second example, the fact that *he* is *known for an album* implies that it is his own album, thus we include this fact in the reference.

| *The school was founded in 1851 as the German-English Academy by a group of Milwaukee German Americans that included educationist Peter Engelmann and hardware wholesaler William Frankfurth* | (The school, was founded by, a group of Milwaukee German Americans)  (a group of Milwaukee German Americans, included, Peter Engelmann)  **(The school, was founded by, Peter Engelmann)** |
| --- | --- |
| *Known for his albums of classical piano compositions with a pop music sensibility, Solo Piano I and Solo Piano II, as well as his MC and electro albums, he is also a producer and songwriter.* | (He, is known for, Solo Piano 1)  (Solo Piano 1, is, his album) |

If the sentence implies that two or more things may be true, but it's impossible to tell which by reading the sentence alone, we don't add this information to the reference. In the example, *The Guardi*an is either a *newspaper* or a *magazine*, and the sentence does not allow to infer which it is. The subject, *He*, has also written at least one *piece* in each publication, but it's impossible to know whether it's a *newspaper* or a *magazine*, or how many he's written, so we omit this information.

| *He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.* | (He, has written [several] magazine opinion pieces in, The Guardian) |
| --- | --- |

**Non-verbal relations**

Some relevant relations are not directly written in the text, even if their information is presented in the sentence. These facts need to be annotated and the words that make up the relationship added, trying to use the simplest, most common relationship that describes the relationship in the most general way possible.

| *1923 Kantō earthquake* | (Kantō earthquake, happened in, 1923) |
| --- | --- |
| *The United Nations (UN)* | (UN, stands for, The United Nations) |

**Reformulation**

If a relation or argument is expressed in a complex way in the text, a simpler re-formulation of the same fact is added in the same cluster, even if the relation in the two formulations is not the same and the level of detail may be different. This is a compromise between the goal of OIE of collecting all the factual information expressed in the text and the importance of formulating these facts in simple language, which is relevant but not necessarily OIE's primary goal. The example shows a case where the reformulated relation is different but conveys the same meaning.

| *Sam managed to convince John* | (Sam, managed to convince, John)  (Sam, convinced, John) |
| --- | --- |
| *Gonzales broadcasts a web series Pop Music Masterclass on WDR, the documentary Classical Connections on BBC Radio 1, The History of Music on Arte, and Music's Cool with Chilly Gonzales on Apple Music's Beats1 radio show.* | (Gonzales, broadcasts Pop Music Masterclass on, WDR)  (Pop Music Masterclass, is broadcasted on, WDR) |

**Active and passive voice**

Clusters are used to group together the active and passive formulations of an extraction in a single fact. If the active formulation is not present in the text, it should still be added in the same cluster if it is simpler than the original formulation present in the sentence. The example shows an originally passive triple and it’s active formulation added in the same cluster.

| *The apple was eaten by Kyle* | (The apple, **was eaten by**, Kyle)  (Kyle, **ate**, the apple) |
| --- | --- |

**Single argument**

Some facts don't have a second argument, so we choose to leave it empty (we annotate XXX for the absent argument for practical purposes).

| *From 1970 to 1985 , Gideon Rodan taught at the University of Connecticut School of Dental Medicine until he switched over to Merck .* | (Gideon Rodan, taught, **XXX**)  (Gideon Rodan, was, a teacher) |
| --- | --- |

**Attribution and speculation**

Some information in the text is speculative or attributed to an entity, so this characteristic must be included in the relationship, in the way it is formulated in the sentence. This makes it possible to preserve this information without having to introduce a particular structure. This information must be included in the relation, as it is in no way related to the arguments. The example shows a case where the attribution is added in the relation of the triple.

| *The earth is flat, according to an Apple Valley man.* | (The earth, is **according to an Apple Valley man** ,flat)  (The earth, is, flat) |
| --- | --- |

**Correction**

Occasionally, some tuples may consist of words from the original sentence but contain grammatical errors. In this case, the tuple formed from the original words and the corrected tuple should be included in the same cluster. This ensures that neither the systems making the correction nor those making it are penalized. The example shows that newspaper without an ‘s’ is a grammatical error, so both the original and the correction should be included.

| *He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.* | (He, has written in, newspaper)  (He, has written in, newspaper**s**) |
| --- | --- |

**Specific information type**

Dates are special types of information, we choose to treat them differently by including only one extraction, as specific as possible. We include the day, month, year, hour, etc. if available. For location-related information, if several specifications exist (Continent, Country, City, etc.) in the sentence, we include one extraction per level of detail.

| *He was born on september 26, 2017* | (He, was born on, **september 26 2017**)  (He, was born in, **september**) |
| --- | --- |
| *He represented the riding of Nickel Belt in the Sudbury , Ontario area .* | (He, represented the riding in, **the Sudbury area**)  (He, represented the riding in, **Ontario**) |