Annotation guidelines for scientific citations in the context of plant health

This guide describes how to establish the following annotation labels on citation passages in the context of plant health, more specifically, in the context of vector-borne plant diseases :

- the rhetorical function: argumentation role of the citation.
- the biological function: biological role associated with the citation.
- the vector status of the citation: does the citation mention an insect as a vector of plant pathogen or a vector associated with a vector-borne disease. If yes, is its exact status of “vector” unambiguous: acquisition and inoculation proven (i.e. vector *stricto sensu*), or not (i.e. potential vector)?
- the polarity: does the author have a judgement about the cited information?

The citation passage is the text segment in which an author addresses a cited work. It is composed of the citation context, the *citance* and the reference. The citance refers to the sentence containing a reference. For the context, we fixed a window of three sentences before the citance and three sentences following the citance, seven sentences in total.

A citation passage can contain several references, in which case we consider them as different citations. An example of this is the following citation extracted from the Pear Decline corpus (detailed corpus description below):

“In most of the insect species tested, we observed high infection rates of CLeu which, in the case of C. pyrisuga, were higher than the percentages observed in Florida, USA (8.7%) and Indonesia (45.2%) for Diaphorina citri with 'Ca. Liberibacter asiaticus' (Manjunath et al. 2008; Subandiyah et al. 2000).”

These two references are related to different information. The first is related to the infection rate of C.Leu (acronym for the bacteria ‘*Candidatus* Liberibacter europaeus’) in the case of C. pyrisuga (insect = psyllid *Cacopsylla pyrisuga*) observed in Florida, whilst the second one is in Indonesia.

“Not surprisingly, some bacteria species utilize SA hydroxylase for a degradation of SA to catechol to suppress an adequate plant defence reaction <ref type="bibr" target="#b65">[66]</ref> or evolved the production of effector proteins that interfere with SA regulated defence responses by activating JA pathway <ref type="bibr" target="#b66">[67]</ref>.”

When establishing the annotations, we first select the citation passage. Then, we determine for each sentence within the citation passage whether it contains useful information to establish the beforemention annotation labels.

The Pear Decline corpus

The Pear Decline corpus is a collection of 74 documents in pdf about the plant pathogen '*Candidatus* Phytoplasma pyri' associated to the disease, Pear decline, created by Myriam Dulor in the context of an internship. The documents were then converted to xml format with GROBID tool[1]. After filtering document types provided by GROBID, we have kept only journal articles which contain primary information. We expect that they would be cited for their finding or claim, and confirmed or debated.

We then filtered relevant journal articles for the topic of insects as vectors of plant pathogens We looked more specifically whether an article was referring to an insect vector. We excluded from this selection articles that had no citation or only one citation, articles that were only discussing the morphology and taxonomy of an insect or that were not mentioning a vector, and articles for which we did not obtain a good quality xml extraction. After this filtering, we consider 24 articles relevant for the topic.

We extracted the citations from the documents using regular expressions. We did not consider citations that contained the pattern "'type="table"'. We also extracted the citation context according to the following window : three sentences on the left and three sentences on the right. We chose this window size based on the work of (Jiang and Chen, 2023) who indicates that the window two left sentence, three right sentence covers most citation context. We decided to add one left sentence to capture further context that could be relevant for identifying vectors and pathogens.

After considering only the citations extracted from those 24 articles, we count a total of 1686 citations. After filtering citations from relevant articles, we randomly chose 100 citations to annotate. We discarded some occurrences such as figures, tables or only dates that have also the tag <ref type="bibr" by mistake. We replaced those with other randomly selected citations.

Citation rhetorical function

The first annotation level is the citation rhetorical function. It is the interpretation of the author's reason for citing (Teufel et al., 2006), aiming to understand the rhetorical role of the citation for the argumentation structure.

Following the work of Simone Teufel (Teufel et al., 2009) and Xiaorui Jiang (Jiang and Chen, 2023) in Computer linguistics, we consider 11 classes for this work: Basis; Contrast/Comparison in Goals or Methods (CoCoGM); Comparison and Contrast in Results (CoCoRes); Comparison and Contrast between cited studies (CoCoXY); Future; Motivation; Neutral; Similar; Support; Usage; and Weakness.

We used those 11 classes as main labels. However, the label "background" was also added as a secondary label for some neutral citations. When the "background" label was added, we

[1] https://github.com/kermitt2/grobid

separated it from the main label with the symbol “|”. The original annotation scheme from Teufel can be found here.

In the following sections, we will define the described annotation labels and types of classes, as well as specific examples from the corpus developed for this work. When no example from our corpus was found, we referenced other examples from biological papers. When further clarification was needed, we added the citation context (in italics) to better understand the citance.

Basis

Definition: The citation acknowledges the intellectual basis of the current work.

Example from our dataset:
“Based on the information obtained in this and previous studies (14,15; A. H. Purcell, unpublished data), pear growers are now aware that pears are the primary reservoir for PYLR in northern California.”

CoCoGM

Definition: Designates a contrast or a comparison in goals or methods between the citing article and the cited article.

Example:
“The PCR amplification conditions were the same as proposed by Ghanim et al . [43] “ (Moreno-Delafuente et al., 2013)

CoCoRes

Definition: Compare or contrast in results. In our annotation, this class is specific to the comparison between the results from the citing article and the cited article.

Example:
“Comparing the detection of grapevine yellows phytoplasma in planthoppers, only 66% of the PCR positives were also positive by enzyme-linked immunosorbent assay (35).

CoCoXY

Definition: Comparison or contrast between two cited studies. In our annotations, CoCoXY can be a comparison or contrast in method as well as in results between two cited articles that are not the citing article.

Example:
“Ullman and Mclean (1986) and Garzo et al. (2012) also observed the same number of teeth

on the mandibles of the psyllids C. pyricola and Diaphorina citri respectively, whereas Pollard (1970) found 8 teeth in adults (7 teeth in nymphs) on the mandibles of C. mali."

Future

Definition: Mentions possible future work.
We did not find this class in our 100 citation sample. Therefore, we will not consider this class for the classification.

Motivation

Definition: Refers to the reasons that justify the current research. These reasons can be promising results or stakes.

Example:
"Regarding the capability of B. nigricornis to transmit CaLsol, previous field work has shown that B. nigricornis can become naturally infected with CaLsol haplotype E (Teresani et al. 2014;2015), so further research to assess the vector efficiency of this psyllid species was needed."

Neutral

Definition: Does not mention cues for classifying the citation passage in the other classes. Often contain background information.

Example:
"The number of protrusions varies among different species, which may be related to the hardness of the leaves of the host plant (Forbes, 1977;Rosell et al., 1995;Zhao et al., Garzo et al., 2012)."

Similar

Definition: Highlights a similarity. (Teufel, 2006) highlights ambiguity between "similar" and "usage" class as an interpretation of the author using the cited method. We did not find this instance in our 100 sample. Despite ambiguity with the similar class, the following example was classified as usage :
"In this paper, times of origin and major divergences of phytoplasmas were estimated using the 16S rRNA gene. This molecular marker has been widely used for phylogenetic and taxonomic classification of prokaryotes (Johansson et al., 1998). "

Support

Definition: Something supports or provides evidence to something.

Example:
"Sequence similarity values within taxa and divergence between taxa largely confirm the results of previous work (Seemüller et al., 1994(Seemüller et al., , 1998b))."

Usage

Definition: Mentions a technique, a tool or data used in the work of the citing article.

Example:
"In direct PCR or the first amplification of semi-nested PCR, the universal phytoplasma primers P1/ P7 (Deng and Hiruki, 1991;Schneider et al., 1995) were used".

Weakness

Definition: Critics or underlines the limits of the cited event. In our dataset, we will designate as weakness the unclear cited statements.

Example:

"To compare our results with those papers, we preferred to use the same methodology despite some potential limitations, such as a poor fit of the K2P model at the species level (Srivathsan and Meier, 2012;Collins et al., 2012).".

Biological function

The biological function refers to the biological information highlighted by a citation. The biological function can refer to one or several objects among: an insect, a pathogen, a plant, a disease, or to an undefined entity. The following classes were established after empirical analysis of a 100 citations sample extracted from a Pear Decline disease corpus.
The general function is separated from the object to which it is associated using the symbol "_".

Location

location_pathogen_vector

Definition: Mentions the location of a pathogen as well as its vector

Example:
"During a survey between 2000 and 2006 Jarausch et al. (2007b) could show that 'Ca. P. prunorum' as well as its vector, C. pruni, were present on all cultivated Prunus species in several stone fruit growing regions in Southwestern Germany."

location_disease

Definition: Refers to a geographic location of a disease

Example: "*ESFY is known to occur in most southern and central European countries (Cieślińska 2011; Marcone et al. 2010, 2011) but it has also been detected in Asia Minor*

(Jarausch et al. 2000; Sertkaya et al. 2005; Tedeschi et al. 2013; Allahverdi et al. 2014; Valasevich and Schneider 2016) as well as in northern Africa (Ben Khalifa et al. 2011).
Its highest spread is in the Mediterranean basin while its northern border ranges from England (Davies and Adams 2000) via Germany (Jarausch et al. 2007b) to Poland."

location_vector

Definition: Mentions the presence of a vector in a location. It also refers to the transmission of a pathogen by a vector in a location.

Example:
"In summary, 27 nominal species of pear psyllids are currently known from China (Li 2011; Luo et al. 2012), three from Japan (Inoue 2010; Inoue et al. 2012), six from Korea (Park 1996;Cho & Lee 2015;Kwon et al. 2016), three from the Russian Far East (Gegechkori & Loginova 1990) and eight valid species from the west Palaearctic, Middle East and Central Asia (Burckhardt & Hodkinson 1986)"

Example of transmission of a pathogen by a vector in a location :
"More recently transmission of the PD agent by C. pyricola in England (Davies et al., 1992) and by C. pyri L. in France (Lemoine, 1991) and Italy"

location_pathogen

Definition: Mentions the presence of a pathogen in a location

Example:
"Liberibacter solanacearum also affects plants from the family Apiaceae such as celery (Apium graveolens L.) and carrot in Europe, Africa and the Middle East (Munyaneza et al. 2010a;2015;Alfaro-Fernández et al. 2012a;2012b;Loiseau et al. 2014;Tahzima et al. 2014;Teresani et al. 2014)."

identification

Definition: Refers to the characterization or identification of an object. This identification is done using different molecular biology techniques such as ELISA. It can also refer to a morphological marker in the case of insects.

identification_unknown

Definition: An identification technique is mentioned but we do not know to which object it is applied.

We created this class for completion, however, we do not have examples from our 100 citation sample. We expect this case to occur if we apply our biological classification to a bigger dataset.

identification_pathogen

Definition: A molecular marker technique is applied to identify a pathogen.

Example:
“Subsequently, several research groups during the late 1980s and early 1990s designed phytoplasma universal (generic) or phytoplasma group-specific oligonucleotide primers that were based on highly conserved 16S rRNA gene sequences”

identification_pathogen | phylogeny_pathogen

Definition: Refers to phylogeny, genetic variation, differentiation of the pathogen genetically, specific type of pathogen compared to other types of pathogen

Example:
“This molecular marker has been widely used for phylogenetic and taxonomic classification of prokaryotes (Johansson et al., 1998).”

identification_insect

Definition: A molecular or morphological marker technique is applied on insects

Example:
“In 1995, groups of 50 individuals of C. pyri were collected from the same infected orchard in July, August and September, respectively, and were then analyzed by PCR. For the analysis, the two different pairs of ribosomal primers used were: fPD/r0 1 (Lorenz et al., 1995) and AP3/ AP5 (Firrao et al., 1994) respectively.”

identification_insect | phylogeny_insect

Definition: Refers to genetic difference or similarity between insects, when there are tests to estimate their genetic variation

Example:
“This has been the most widely used method for DNA barcoding analyses (...). To compare our results with those papers, we preferred to use the same methodology despite some potential limitations, such as a poor fit of the K2P model at the species level (Srivathsan and Meier, 2012;Collins et al., 2012). *Eleven pear psyllid species of the genus Cacopsylla (Psyllidae: Psyllinae) were included into the analyses and two Acizzia species (Psyllidae: Acizzinae) were used as outgroups (table 1)*.”

relation

We can distinguish three different types of relation between the objects: the biological interaction, the association (x disease is associated to y insect), or an interaction wrongly designated (it should be an association but the author describes it as an interaction)

The different objects of the relation are separated by “_”.

relation_insect_disease_plant

Definition: mentions a relation between an insect, a disease and a plant, but not the pathogen

Example:
“Bactericera cockerelli was associated with zebra chip disease in potato in 2007 (Munyaneza et al. 2007) and later with CaLsol in 2009 (Secor et al. 2009)”

relation_insect_disease_pathogen

Definition: mentions a relation between an insect a disease and a pathogen, but not the plant

Example:
“After 1970, mycoplasmalike bodies were detected in sieve tubes of pear trees affected with pear decline (Hibino and Schneider, 1970) and in the pear psylla vector of pear decline.”

relation_insect_pathogen_plant

Definition: mentions the biologic interaction between an insect a pathogen and a plant , but not the disease

Example:
“By contrast, B. nigricornis is known to be able to properly feed from the phloem and colonize potato crops (Fathi et al. 2011; Antolínez et al. 2019), which facilitates the transmission of phloem-restricted pathogens.

relation_insect_plant

Definition: mentions a relation between an insect and a plant without mentioning the pathogen, or the disease

Example:
“Although eggs and immatures B. nigricornis are rarely observed during visual inspections in the field, the reproduction and presence of this psyllid species on potato crops has been confirmed (Hodkinson et al. 1981; Antolínez et al. 2019) and has been reported to cause severe yield losses in Iran (Fathi et al. 2011).

relation_insect_pathogen

Definition: mentions a relation between an insect and a pathogen without mentioning the plant or the disease

Example:
“Regarding the capability of B. nigricornis to transmit CaLsol, previous field work has shown that B. nigricornis can become naturally infected with CaLsol haplotype E (Teresani et al. 2014; 2015), so further research to assess the vector efficiency of this psyllid species was needed.

relation_pathogen_plant

Definition: mentions a relation between a pathogen and a plant

Example:
“A previous report detected WX in pear by enzyme-linked immunosorbent assay (ELISA) (9), but these results were never repeated using DNA hybridization (B. C. Kirkpatrick, unpublished results) or PCR (5)”

plant-modification-by-pathogen

Definition: It is the consequence of the relation pathogen plant. It refers to the effects of the pathogen on the plant : how the affected plant reacts to a pathogen : symptoms, defenses, hormonal changes…

Example:
“Not surprisingly, some bacteria species utilize SA hydroxylase for a degradation of SA to catechol to suppress an adequate plant defence reaction [66] or evolved the production of effector proteins that interfere with SA regulated defence responses by activating JA pathway [67].“

life-cycle_insect

Definition: Refers to the seasonal changes and to the different instars of an insect. It includes the “remigrant” notion

Example:
“At 15 to 33 °C, female longevity varies from 88.3 to 28.7 d and is 38 % higher than male longevity (Liu & Tsai 2000;Nava et al. 2007).”

infection-rate

Definition: mentions the infection rate of a pathogen transmitted by an insect or the number of infected trees after inoculation experiments. We include in this class cases when :

- the infection rate of a pathogen is transmitted by an insect in a specific location
- the infection rate of a pathogen is transmitted by an insect in specific climatic or seasonal conditions

Example:
“Apple trees were infected with a virulent accession (3/6, n = 13) in 2017 [77] [78] [79].”

sampling

Refers to the processing of collecting samples, they can be insects or pathogens

sampling_insect

Definition: methodology on how the insects were collected for the experiment

Example:
“The psyllids were then collected in glass tubes by means of a mouth aspirator and brought to the laboratory for further analyses (Horton 1994; Tedeschi et al. 2002).”

sampling_pathogen

Definition: methodology on how the pathogens were collected for the experiment

Example:
“We contributed 22 German isolates covering 5 stone fruit growing regions with 16 apricot or peach orchards to the MLST analysis of Danet et al. (2011).

other

insect-management

Definition: refers to methods used to regulate insects (vectors or pests)

Example:
“As facultative endosymbionts can render insects less susceptible to management using parasitoids, pathogens, host plant resistance, and insecticides (Oliver et al. 2003, Scarborough et al. 2005, Hansen et al. 2007, Ghanim and Kontsedalov 2009, Kikuchi et al. 2012, Su et al. 2015), the documentation of the presence and biological function of endosymbionts in pest populations can allow for more informed pest management decisions.”

disease-management

Definition: refers to methods used to treat and regulate a plant disease

Example:
“Subsequently, a few additional divergences of major phytoplasma lineages occurred between the Carboniferous and the Cretaceous but most of the recognized modern 16Sr phytoplasma groups did not appear until later in the Cretaceous or during the Cenozoic, after the radiation of angiosperms (168-246 Ma;Kumar et al., 2017;Morris et al., 2018)”

disease-impact

Definition: mentions plant disease and vectors impact on economy or agriculture

Example:
“Their economic importance has risen in the past 20 years probably due to globalization (increasing trade of plant material over the world) and also due to global warming that facilitates the expansion and adaptation of psyllid pests into new habitats and geographical regions (Fereres 2015).

behaviour

Refers to behaviour or adaptation of the insect or the pathogen.

behaviour_insect

Definition: refers to behaviour or adaptation of an insect

Example:
“Psyllid nymphs and adults feed on plant phloem and occasionally on xylem sap [31] [32][33].”

behaviour_pathogen

Definition: refers to behaviour or adaptation of a pathogen

Example:
“Tenericutes bacteria have evolved a broad range of lifestyles, including free-living, commensalism and parasitism (Razin, 2006; Rivera-Tapia et al., 2002; Tully, 1996).”

insect-morphology | insect-anatomy

Definition: refers to the anatomy / morphology of an insect

Example:
“The adult mouthpart morphology of C. chinensis is generally similar to that of other sternorrhynchan species described previously (Pollard, 1973; Tavella and Arzone, 1993; Rosell et al., 1995; Boyd, 2003; Wiesenborn, 2004; Rani and Madhavendra, 2005; Anderson et al., 2006;Garzo et al., 2012), consisting of a three-segmented labium which lies between the prothoracic coxae of the first and second pair of legs with a deep groove in the anterior side, a moderate number of sensilla, a stylet fascicle consisting of two mandibular and two maxillary stylets which are conjoined together forming a food canal (Fc) and a salivary canal (Sc), and a triangular labrum.”

pathogen_transmission

Definition: mentions a pathogen transmission process

Example:
“Candidatus Liberibacter solanacearum is an intracellular phloem-limited bacterium that infects and replicates in the sieve elements of plants as well as in different organs and tissues of its psyllid vectors (Brown 2016; Haapalainen 2014;Perilla-Henao & Casteel 2016).”

plant_biology

Definition: refers to plant biology such as phloem structure

Example:
“Liu and Gao (1993) found similar phloem structures in Malus and Pyrus in comparison to Prunus, too [50].”

Vector class

not_in_citation

Definition: The citation does not mention a specific vector or does not mention vectors at all

Example:
“In addition to the adverse ecological effect of declining vector abundance, vector management measures, such as insecticides, might impact ecosystem functioning by directly harming nontarget species; this is well illustrated in studies of tsetse flies and ticks, in which effects on nontarget species of insecticide-impregnated traps/targets and 'pour-on' for cattle (a mixture of repellents and insecticides) destabilize food webs, with cascading adverse effects on biodiversity [94]. *Similarly, for agriculture, the resurgence of pest outbreaks or epidemics can often be associated with a breakdown in multitrophic relationships due to the unintended effects of insecticides on nontarget organisms*”

Y

The citation mentions a vector:

- **confirmed** : the author confirms that the insect is a vector
  We did not find this occurrence in our 100 citations sample. We expect to find this case if we apply our analysis to a bigger dataset.

  Example from (Koudamiloro et al., 2015) : “T. sericea was confirmed as vector in Niger [15], Mali [60], Cameroon [51, 61], and Ivory Coast [6, 52, 59].”

- **factual** : the author mentions a citation which report an insect as a vector

  Example:
  “More recently transmission of the PD agent by C. pyricola in England (Davies et al., 1992) and by C. pyri L. in France (Lemoine, 1991) and Italy (Carraro et al., 1998a) , has been reported.”

- **unconfirmed** : citation mentions insect as a potential vector for which further studies are required to prove transmission or if the fact that the insect is a vector has never been proved

  Example:
  “Regarding the capability of B. nigricornis to transmit CaLsol, previous field work has shown that B. nigricornis can become naturally infected with CaLsol haplotype E (Teresani et al. 2014;2015), so further research to assess the vector efficiency of this psyllid species was needed.”

- **ambiguous** : the citation does not state clearly whether the insect is a vector but seems to suggest it

Example:
“Bactericera cockerelli was associated with zebra chip disease in potato in 2007” (Munyaneza et al. 2007) and later with CaLsol in 2009 (Secor et al. 2009).”

insect_mentioned

Definition: mentions the presence of an insect in a plant but we don’t know if it is a vector

Example:
“From the Russian Far East, finally, three species associated with pear are reported to date (Gegechkori & Loginova 1990): the west Palaearctic C. pyricola and C. pyrisuga, and the native Psylla nigrella Konovalova.”

Polarity

neutral

Definition: The citance does not have a judgement about the cited information

Example :
“The lower temperature threshold for development is about 10 °C (Catling 1973), whereas a temperature of 32 °C, with low relative humidity, is particularly deleterious for all instars (Moran & Blowers 1967)”

positive

Definition: The author agrees with the cited information

Example :
“Sequence similarity values within taxa and divergence between taxa largely confirm the results of previous work (Seemüller et al., 1994 (Seemuüller et al., 1998b)).”

negative

Definition: The author has doubts, does not agree or highlights limitations about the cited information

Example:
“Several previous authors have accepted the idea that the Cacopsylla species present from the region is Cacopsylla pyricola (Forster, 1848)(Ahmed and Ahmed 2013;Abrol 2015;Mahendiran et al. 2016. *However, there are no reliable distributional records of this species from India (Ouvrard 2017)*.”

Reference

Jiang, X., & Chen, J. (2023). Contextualised segment-wise citation function classification. Scientometrics, 128(9), 5117–5158. https://doi.org/10.1007/s11192-023-04778-3

Moreno-Delafuente, A., Garzo, E., Moreno, A., & Fereres, A. (2013). A plant virus manipulates the behavior of its whitefly vector to enhance its transmission efficiency and spread. PLoS ONE, 8(4), e61543. https://doi.org/10.1371/journal.pone.0061543

Koudamiloro, A., Nwilene, F., Togola, A., & Akogbeto, M. (2015). Insect vectors of rice yellow mottle virus. Journal of Insects, 2015, 721751. https://doi.org/10.1155/2015/721751

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 103–110.

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). An annotation scheme for citation function. Proceedings of the SIGDIAL Workshop on Discourse and Dialogue, 80–87.