Instructions for Labeling Sentences

These instructions describe how to label sentences in scientific articles. The labels are:

Aim	AIMX	The specific research goal of the paper
Own	OWNX	The author's own work, e.g. methods, results, conclusions
Contrast	CONT	Contrast, comparison or critique of past work
Basis	BASE	Past work that provides the basis for the work in the article.
Misc	MISC	Any other sentences

The labeling procedure

Step 1: Before labeling

*Skim read the paper **before** you start labeling*. Trust me, this really helps!

Try to get a feel for the overall flow and structure of the paper. Does the author begin with general background information? When do they transition to discussing their own work?

Don't try to understand all of the scientific details.

Step 2: Labeling

Labeling proceeds sentence by sentence. Each sentence can only have one label.

The process you should follow to label sentences is shown in Figure 1. For each sentence, start at the top of the decision tree. Answer each yes-no question and follow the arrows until you reach a label. If there is more than one label that fits, use your judgment to pick the best label.

Treat the sentences in the way the author intended. Often, the context and location of a sentence are important clues.

You can label consecutive sentences with the same label if they *together* fulfill the criteria of the label. E.g. you could mark two sentences as AIMX if they together describe the specific goal of a paper well.

All citations have been replaced with the marker "CITATION". All numbers have been replaced with the marker "NUMBER". All equations and symbols have been replaced with the marker "SYMBOL".

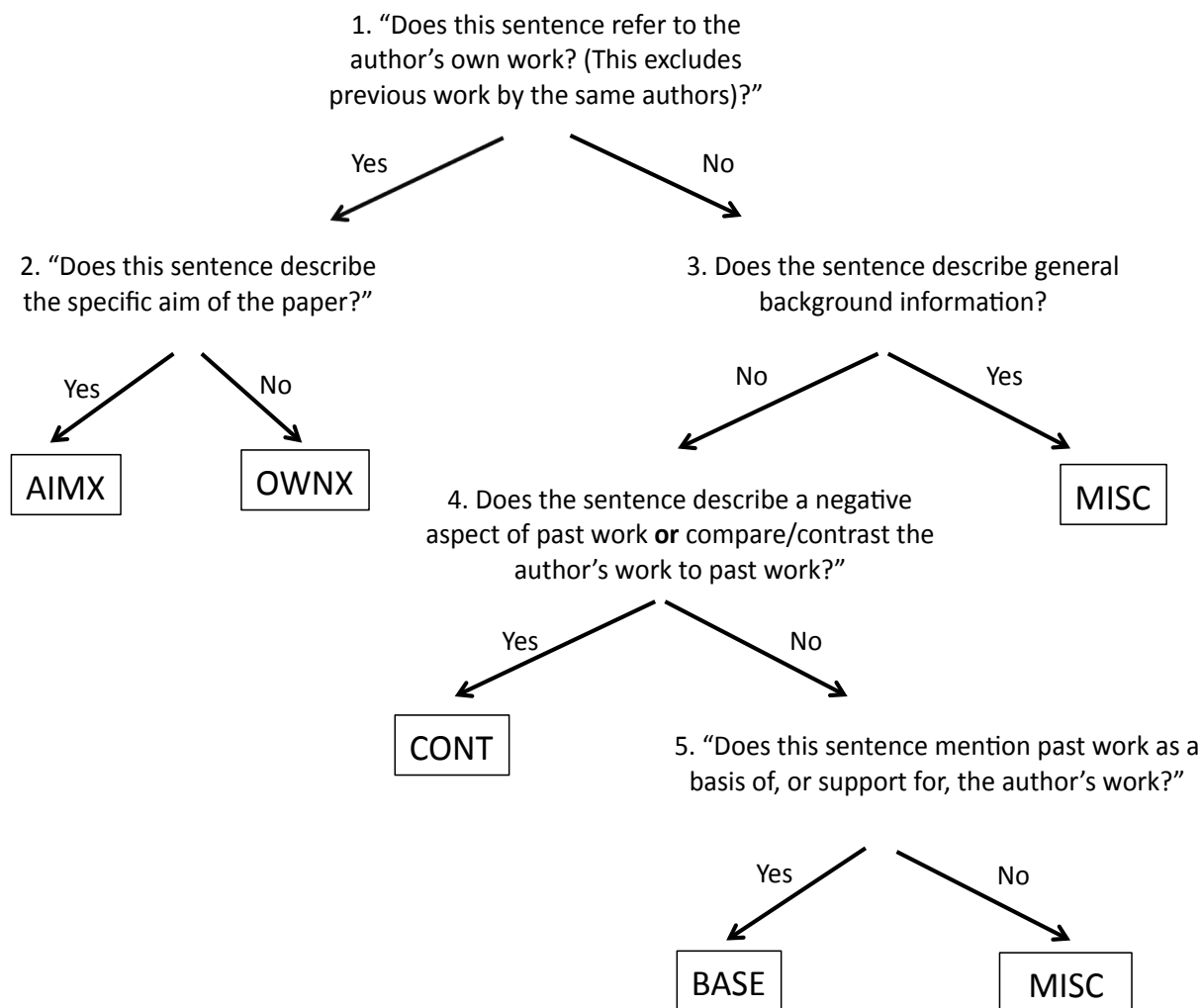It is possible that the article will not have any CONT or BASE sentences.

Figure 1: Use this decision tree to label each sentence.

Step 3: After labeling

Check a few things, and fix your labels if necessary:

1. There should almost always be an AIMX sentence in the abstract and an AIMX sentence in the introduction for any well-written article. If you did not mark an AIMX sentence in both sections, ***strongly consider re-labeling sentences*** until you have at least one AIMX sentence in both sections.

The Questions

Question 1: Does this sentence refer to the author's own work? (This excludes previous work by the same author)?

This question is meant to distinguish between the author's own work and all other work.

The author's own work refers to the work that was performed by the author and is being presented in the current article, e.g. the methods, results, conclusion of the article. This does not include previous work by the same author.

This includes both **direct statements** of the author's methods, results, conclusions, etc. as well as **descriptions**, and **elaborations**, of the author's methods, results, conclusions, etc.

Often, these sentences contain explicit references to the author using words like "we", "our", "I", or "the authors".

It is common to see an OWNX sentence that contains one of these words, e.g. "our", followed by one or more OWNX sentences that do not contain any of these words but elaborate upon the first sentence. For example,

> "Our program calculates likelihood scores. The program then determines the pairs with the greatest local likelihood."

The first sentence has an explicit reference to the authors "Our program". The second sentence, which describes the model, does not have an explicit reference to the authors but it is still referring to the author's own work. Both of these sentences should be labeled OWNX.

Question 2: Does this sentence describe the specific aim of the paper?

Two kinds of sentences count as aim statements:
- Goal statements (i.e. description of research goal)
- Scope statements (i.e. delimitation of research goal: what the goal is not)

If the sentence describes a general goal in the field, it should **not** be marked as AIMX. For example,

> "Constraining free parameters in mathematical models is a major challenge in computational biology"

This describes a goal of computational biology – not of the paper. AIMX sentences describe particular goals of the paper. There are different ways of expressing the particular goal of the paper.

Directly: Directly state the aim or goal of the paper.

- "We aim to find a method for…"
- Our goal, however, is to develop a mechanism for…"
- "I show how X can be induced."
- "In this paper, we address two issues relating to…"

Functionally: Another way of expressing a research goal is to say that one has accomplished doing a certain task.

- "We have presented an analysis of X"
- "This paper gives an algorithm for doing X..."
- "We have improved the accuracy of the software by..."
- "We have developed a method that combines these two approaches..."

Hypothesis: The aim might be expressed as a hypothesis.

- "We tested the hypothesis that..."

Focus: The aim might be presented as the focus of the paper.

- "In the present study, we focused on one aspect of..."
- "We used a computational approach to investigate...with specific focus on..."
- "This paper concerns the formal definitions of grammars"

Scope: These sentences define the goal as *part* of a previous goal, excluding some other similar goals.

- "Here we will look at only relative pronouns".

"You must be sure that the [sentence you mark as AIMX] is indeed important enough to cover the goal of the whole paper, and you must be sure that the sentence refers to the *highest* level of problem solving. If it talks about [or addresses] a *sub*problem, don't consider the sentence an AIMX"

If you answer "yes" to Question 2, then label the sentence as AIMX.
If you answer "no" to Question 2, then label the sentence as OWNX.

Question 3: Does the sentence describe general background information?

Background sentences are sentences that are presented as **uncontroversial** in the field or **undisputed** – i.e. knowledge that is generally **accepted by everyone**.

These sentences help establish the context of the author's own work.

These sentences are often in the beginning of the abstract or the introduction. However, sometimes an author starts talking immediately about their own work and skips any background information.

Some types of background sentences include:
- Statements of what is possible in a particular field.
- Challenging problems in a field
- General solutions to problems
- Facts
- References to pioneers in a field

Example sentences include:

- "In recent years, there has been a growing interest in the field of X…"
- "One of the difficult problems in the field of biology is…"
- "High mutation rates of HIV confer upon HIV resistance to drugs."
- "Tagging by means of a Hidden Markov Model (HMM) is widely recognized as an effective technique"

If you answer "yes" to Question 3, mark the sentence as MISC

When to answer "no" to Question 3?

In general, if you are unsure whether a sentence describes background information, answer "no" to Question 3.

You should also answer "no" when the sentence describes past work specifically enough that the author could contrast or compare their work to it.

The past work must be *clearly attributable* to some other researchers – otherwise it might be too general, in which case, you should consider it background.

Common ways of attributing research is to use a citation, or to include the names of the researchers, or to give a name to a solution or method. For example,

- "Johnson et al. argue that children don't acquire grammar frames until…"
- "The bilingual dual-coding theory CITATION partially answers the above questions"
- "Similarity-based models suggest an appealing approach"

These sentences are not background sentences. They describe past work specifically enough that the author could critique them.

"Stop marking sentences as MISC when you reach a point where ideas, solutions, or tasks are…attributed to researchers in such a way that they can get criticized. "

Question 4: "Does the sentence describe a negative aspect of past work or compare/contrast the author's work to past work?"

Other past work is described in a contrastive way, if the sentence describes:
- **Weaknesses** of other people's solutions
- The **absence of a solution** for a given problem
- **Difference in approach**/solution
- **Superiority** of own solution
- Statements of **direct comparison** with other work or between several other approaches (these appear mostly in evaluation papers)
- **Incompatibility** between own and other claims or results

Read through this list carefully so you remember the types of contrast sentences that are possible!

Note that sentences which state the absence of a solution to a problem should be marked as CONT.

Here are some examples of CONT sentences:

- "X's solution is problematic for several reasons."
- "A limiting factor if this model is…"
- "Although much work has been done on X, little attention has been paid to Y"
- "In contrast to standard approaches, we use a statistical model"
- "Our model outperforms simple matching by 25%"
- "This result challenges the claims of recent theories"
- "No models exist that describe HIV dynamics under the influence of mutations…"

If you answer yes to Question 4, then label the sentence as CONT.

Question 5: "Does this sentence mention past work as a basis of, or support for, the author's work?"

These sentences state that some other past work is the basis for the author's own work in the paper.

Mark sentences which contain:
- Explicit statements that the author's solution **builds on** the other past work
- Statements that the authors have **adapted** a solution, contributed by someone else.
- Statements about **consistency** with another theoretical framework or other people's results
- Statements of **similarities** between the author's work and the other approaches if these similarities are not cancelled later by mentioning a contrasting property
- Statements of the quality of the other approach

Examples include:

- "We base our model on Y's model"
- "Our approach is in the spirit of Y's approach"
- "We choose to use XYZ-algorithm CITATION"
- "We extend the model for doing X by allowing it to do Y too"
- "This model draws upon CITATION, but adapts it to…"
- "Our account…fits within a general framework for…"

- "The analysis presented here has strong similarities to analyses of the same phenomenon discussed by CITATION and CITATION"
- "We discuss the advantages of CITATION's model"

If you answer "yes" to Question 5, label the sentence as BASE.

If you answer "no" to Question 5, label the sentence as MISC.

After you have finished reading these instructions, please email the time of day to annotation.study@ics.uci.edu so we can confirm that you read the instructions.

These guidelines, instructions, labels, and any quotes have been taken (and modified) from the PhD thesis of Simone Teufel.