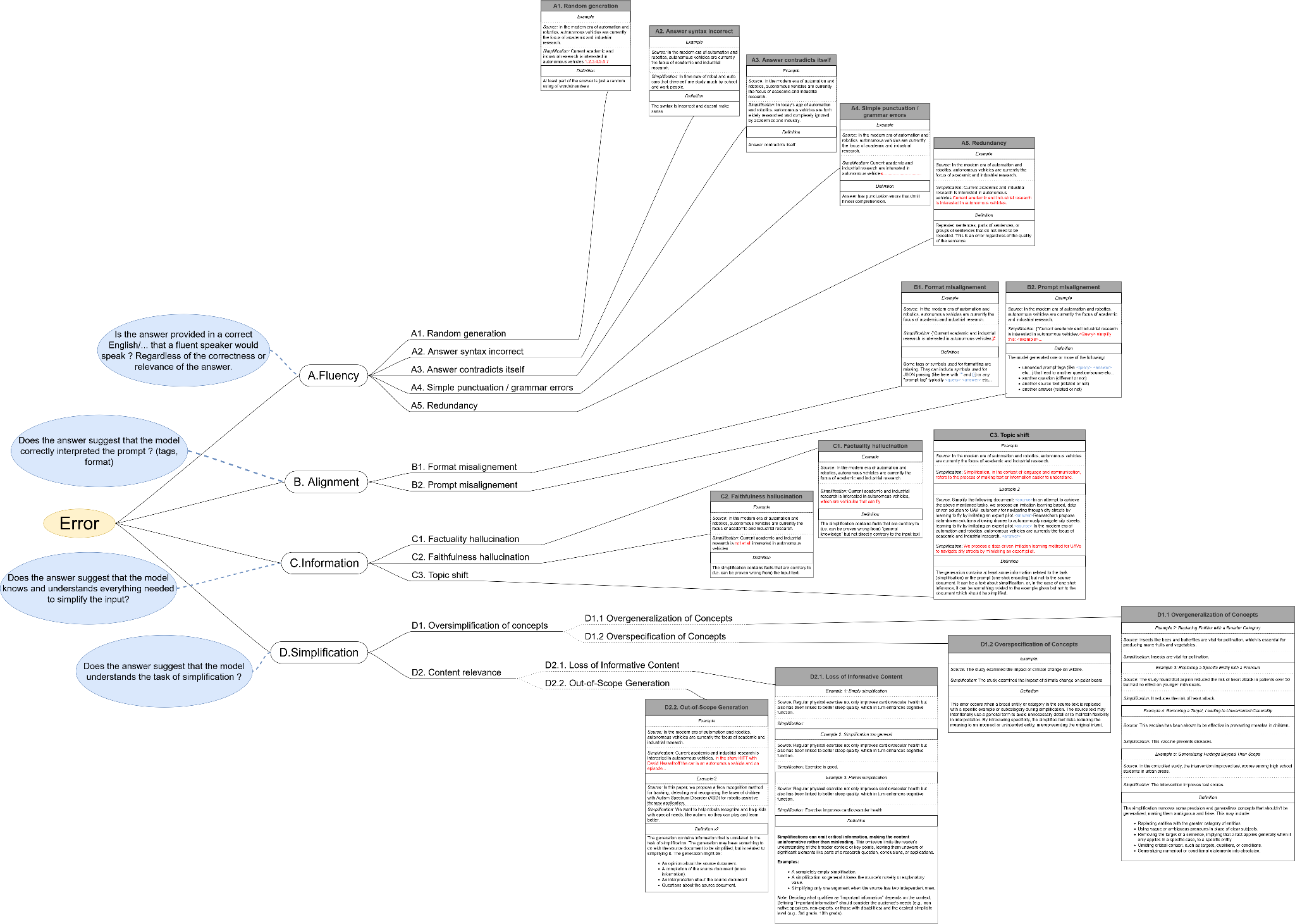
Annotation scheme

On errors in automatic text simplification

The annotation scheme is based on this typology, you can find it in higher resolution in the folder: [tree typology.png](https://drive.google.com/open?id=1EArDWfFd04pmG0n_-Vn-fih_O8PhuTfn&usp=drive_copy)



# Problem:

Automatic text simplification often results in erroneous outputs, referred to as "generations." By "erroneous generations," we do not simply mean outputs that fail to achieve good simplifications, but rather outputs that fail to meet the standards of syntactic or lexical simplification. For example, what we might consider “not good simplifications” are:

* Some words are too technical and should be simplified more
* Sentences are too long
* Sentence syntax is too complex
* Some definition might be needed but is not provided
* Some theory is used by name but not explained
* …

These issues vary depending on the specific goals and requirements of the simplification task. Factors such as the intended audience (e.g., non-experts, non-native speakers, individuals with disabilities) and the desired level of simplification (e.g., high-to-low complexity, middle-to-high school readability) influence what constitutes a "good simplification."

For the purpose of this study, however, we are not concerned with defining or achieving these simplification goals. Instead, our focus is on identifying generations that fail fundamentally from a linguistic, Natural Language Processing (NLP), or data integrity perspective. Meaning, they either aren't proper language (linguistic), don't understand the task (NLP) or the information to simplify (data integrity). These are outputs so flawed that they reflect not just a model’s inability to produce an *effective* simplification, but an inability to produce a *valid* simplification at all. To address this, such errors need to be rigorously defined.

# Goal:

Our goal is to develop a typology representative of errors that occur in automatic text simplification. More specifically, we focus on automatic simplification of scientific texts for non-experts. This typology needs to be precise enough to allow for an easy annotation.

We then want to annotate automatic simplifications in order to:

* See which errors appear more often
* See if some errors appear more in some contexts (source theme, length…) than others
* See if some models are more susceptible to some errors than others
* See if error detection models work better on some errors than others

These insights will guide efforts to improve error detection methods within the context of scientific text simplification for non-experts.

# Typology:

The typology is defined as a tree. Here we will go into the ideas of the structure of the tree, the examples and definitions of the leaves of the tree are available on the graph.

First, we first have three “*Greater types of errors*”.

## Greater Types of Errors

We can define 3 greater type of errors, related to where that error comes from, what made the model make the error:

* A. Fluency:
  + Is the answer provided in a correct English/... that a fluent speaker would speak ? Regardless of the correctness or relevance of the answer.
  + This category focuses on the model's knowledge of the language itself, which is typically acquired during its initial training phase. Not learning it properly can lead to errors in this category.
* B. Alignment:
  + Does the answer suggest that the model correctly interpreted The prompt ? (tags, format)
  + This category relates to the model's ability to follow instructions. It includes learning how to interpret prompts, including one-shot or few-shot examples, and adhering to any specific tags or formats required in practical applications. A bad alignment results in errors in this category
* C. Information
  + Does the answer suggest that the model knows and understands everything needed to simplify the input?
  + This category is related to how well the model handles the information included in and related to the source text. Simplifying text requires a proper understanding of the source material, its concepts, and definitions, but also of the topic as a whole. The simplification must also preserve the accuracy of the original information without introducing distortions or misrepresentations. Simplification that do not respect the information and topic of the source document, result in errors in this category
* D. Simplification:
  + Does the answer suggest that the model understands the task of simplification ?
  + This category is related to the understanding of the task, of the processes related to simplification. Simplifying text might require removing some unimportant information, but both the amount of removal and definition of “important information” are defined by the understanding of the task. Likewise, simplifying means some concepts should be generalized (“dogs” → ”animals”) or specified (“dogs” → ”animals”) which may lead to a text that is not informative enough, or plain wrong.
  + Different target audiences have different simplification needs. Some level of un informativeness can be accepted, and specific transformations might be suitable for one type of simplification but not for another. This is the category that is subject to change depending on the simplification requirements.

Then, the tree defines actual specific errors. Some of them are grouped in groups of errors.

## Multi Label Annotation

Finally, it is important to remember that this is a “multi-label” process, meaning that ***generations can have multiple errors at once***. <- important

What we mean by that is that sometimes a generation contains multiple errors. In that case, the simplification should have the labels corresponding to each error.

Example:

* *Source sentence:*
  + After an initial implementation period, we performed a pilot study to test the use of this system.
* *Simplification:*
  + <Answer> we performed a pilot study to test the use of this system.,, we performed a pilot study to test the use of this system 13.14.5.15.3.2.2.45.45645123128490.45912.54612

In that example, we have a repeat prompt (*B1.2 Prompt misalignement*), a repetition of words(*A1. Random generation*) and points (*A4. Simple punctuation / grammar errors*)

Here this simplification should have the three labels: *B1.2*, *A1*, *A4*.

In that case, you should check each corresponding box.

# Methodology:

Before annotating, it is essential to read this document and review the typology tree. The tree should be read left-to-right and then top-down, starting with the broader categories of errors and moving to the detailed typology of errors. Similarly, for alignment/prompt and information errors, follow the same left-to-right, top-down approach. For each error type, make sure to read both the definitions and the examples provided to fully understand them.

You should look at annotating in that order:

1. *Fluency*: Is the generation proper English that doesn't repeat or contradicts itself ?
   1. *Yes* → No Fluency error.
   2. *No* → Annotate the proper error. If it is random generation or meaningless syntax, no need to look at other errors.
2. *Prompt*: Does the generation contain only the one simplification, no tags or other question ? And are the eventual “ or { properly opened and closed ?
   1. *Yes* → No Prompt error.
   2. *No* → Annotate the proper error.
3. *Information*: Is the generation on the same topic as the source and doesn’t contradict it or the general knowledge ?
   1. *Yes* → No Information error.
   2. *No* → Annotate the proper error. If it completely changes the topic, then no need to look at Simplification errors.
4. *Simplification*: Is the generation really a simplification ? Do we keep enough important information without adding unnecessary information ? Does the simplification of the concepts lead to misleading or out-right false information ?
   1. *Yes* → No Simplification error.
   2. *No* → Annotate the proper error.

## Annotation Process

You will be provided with a Google Sheet personalized with your name (an example dataset is also available in the shared folder for reference).

* **Structure of the Sheet:**
  + On the left, you will find the source sentence alongside its corresponding simplified sentences.
  + The columns run\_id and snt\_id are hidden.
  + On the right, you will see columns representing the various error types. These columns are color-coded to reflect their broader error categories.
  + For ease of use, each error cell includes a note with the error definition. However, if any clarification is needed, refer back to the typology tree for definitions and examples.
* **Marking Errors:**
  + For each row, tick the boxes corresponding to the identified errors.
  + If no errors are found in a row, tick the "No Error" column. This serves two purposes:
    1. Ensures no rows are accidentally skipped during annotation.
    2. Helps identify where an annotator stopped, should they be unable to complete the full dataset.

## Comments

Sometimes it may be hard to say if something qualifies as an error. In that case, annotators can use the “Comment” column to explain in a few words why this is ambiguous and their doubts. However, this column should not be used as default and rather in a few edge cases. Annotators are advised to trust themselves and their understanding of the typology as much as possible.

## Fluency Errors

The categories under **Fluency Errors** might sometimes appear ambiguous. While fluency-related errors can be noticeable. Annotators are advised not to spend too much time deliberating over what technically qualifies as a syntax error.

The focus of this study lies primarily on the other error types, so it's acceptable to allow for some imprecision in fluency annotations to achieve greater accuracy in the other categories. This approach is based on the following considerations:

* Modern models generally handle fluency well, and a reasonable level of fluency is expected by default.
* Fluency has already been extensively studied, particularly in fields like Machine Translation, and replicating this work is not the objective of the current study.

# Organization:

I need as much data as possible. Right now I have assigned you around 50 rows per person, but I can generate more. As soon as you finish, you can ask me for more. If possible, I would like to have an idea of how much time it takes you to annotate these, but it’s not really important, don’t overstress it.