**Viktoria Csink**

# CheMastery: Identification of relevant information in chemistry recipes

## Task objective

The purpose of this task was to identify specific pieces of information related to the addition of chemical entities during a chemical experimental test. The ultimate goal of the work was to gather the information into a tidy, reliable, machine-readable format in order to guarantee the success of future experiments and to ensure replicability.

The exercise consisted of three tasks: (i) identifying the type of addition (in portions or continuous), (ii) identifying the chemical components that were added and (iii) extracting the quantity of the ingredients used. In the following, I outline my strategy of answering these questions in the order of presentation in my script (Chemistry_recipes.ipynb). The results are displayed in preprocessed_data/Recipes_preprocessed.json.

## Task 1. Ingredients

I first decided to identify the chemical ingredients used throughout the recipes. I reasoned that any chemical that is mentioned in the sentences describing the addition process will most likely be the ingredients of the recipe.

Therefore, I first extracted the sentences in the text using the ChemDataExtractor package up until the last mention of the phrase ‘added’ or ‘addition’. Subsequently, I used ChemDataExtractor to identify named entities in the resulting sentences in each recipe, which derived the ingredients of that recipe.

In addition to this, I reasoned that during the execution of the recipe, the order of adding the various components to the mixture is also an important piece of information (i.e. ‘what is being added to what’). In order to extract this piece of information, I used the *displacy* library in *spacy* to extract and visualise the sentence structure of the recipes. In passive voice (‘X was added to Y’), X is the subject denoting the entity being added, and Y is the object, referring to the entity that is already present in the mixture. Consequently, I extracted the dependency of each token in the sentences describing addition, and identified the named entities in the object position

as “recipients” and the named entities in the subject position as materials “added to the mixture”. In case the named entity was neither in the position of the subject, nor in the position of the object, the results return the phrase “order of addition unknown”.

**Output**: ***Ingredients (e.g. “4-methylmorpholine n-oxide, Order of addition: Added to mixture.”).***

**Task 2. Quantities**

In order to extract the quantities of the chemical entities, I followed two approaches. First, I noticed that certain quantities were displayed in brackets after the named entity (e.g. *potassium osmate dihydrate (97.3 mg, 0.38 mmol)*). Second, some information regarding quantity was presented shortly before the occurrence of the named entity (e.g. *2 ml dry toluene*).

Therefore, in my first approach I constructed a regular expression that captures any pattern that consists of a named entity, immediately followed by the information in brackets. If this search failed to return any results, the code tokenizes the sentences and returns the 3 words immediately preceding the named entity, thus capturing the quantity information before the name.

**Output**: ***Quantities (e.g. 2-methylbenzaldehyde (1.23 g, 10.2 mmol, 2.5 equiv, 97%); 1 ml dry methanol ).***

**Task 3. Addition in portions or continuous**

I approached this task in a data-driven way and constructed two lists which contain key phrases that describe continuous and in-portion addition, respectively. I identified the following phrases as indicative of continuous addition: *by syringe, slowly, dropwise, with stirring*. In contrast, the phrase *single portion* was interpreted as indicative of in-portion addition. (I added the phrases through two txt files with the phrases, in case the future users of the scripts decided to amend these lists.)

Consequently, the type of addition will be determined by the presence of either continuous or in-portion key phrases. Recipes that contain both types of information (N = 1) are tagged with the phrase “mention of both continuous and in-portion

addition". Recipes that do not contain key phrases for either type of addition are returned with the missing information indicated in the results.

**Output**: ***Type of addition: Continuous/In-portion/Mention of both continuous and in-portion addition/Unknown***.

## Limitations and future work

Future work should investigate further how to determine the order in which the chemical components were added to the mixture. While identifying the subject and the object might be a good starting point, the syntactic position of the 'recipient' and the 'new component' are different in the active compared to the passive sentences. Therefore, the approach that worked with the sentence structure "X was added to Y" will not work with the structure "I added X to Y". Secondly, addition was at times expressed with the phrase "the addition of X", which introduces a different grammatical structure.

Similarly, identifying the type of addition (in portions/continuous) needs further work, either via incorporating a large number of relevant phrases, or by looking at the sentence structure of the recipe. The phrase that describes the type of addition is likely to be an adverb in the verb phrase 'add' (e.g. '*X was added dropwise*'), therefore returning the word in this position might capture the type of addition used. However, other, irrelevant words might also stand in this position which do not describe the type of addition.

As a long-term project, I believe that machine learning models could also be implemented to predict which recipes are ill-formed and therefore need expert attention, and which ones are likely to succeed.