Preregistration:

1) Have any data been collected for this study already?

    - We will use an open source corpus (projet Orfeo?)

2) What's the main question being asked or hypothesis being tested in this study?

    - What factors influence the "se faire" vs. "être" passive alternation in French?
    - In other words: what factors influence speakers' choice between the "se faire" and the "être" passive construction?
    - Operationalized as: the probability that the "se faire" passive construction is used

3) Describe the key dependent variable(s) specifying how they will be measured.

    - DV: type of passive construction ("se faire"/"être")
    - Values: 0 = "être" construction; 1 = "se faire" construction
    - Criteria: … (see 6))

4) How many and which conditions will participants be assigned to?

    - This is a corpus study, so there will be no participants.
    - Factors for the logistic regression:
        - Subject
            - Animacy (inanimate/animate)
            - Person
            - Number
            - Gender?
        - Verb
            - Aspect/ Form of auxiliary verb ? (…)
            - Tense ? (…)
            - Main verb semantic group (clusters): (e.g. dynamic/ stative, cognition, movement, perception…)
            - Telicity of the verb (telic/atelic)
            - Adversativity of the verb → sentiment analysis (not adversative/adversative) and (neutral/not neutral)
        - [Modality? (spoken/ written) → probably not, see 5)]
        - [Register? (informal/formal) → probably not]
        - Complément d'agent (absent/present) (if enough data)
        - [Periphrase constructions (present/absent) → probably too few in data]
        - [Negation (present/absent) → probably not, see 6)]

- Interactions tested:
  - Animacy of subject and Complément d'agent ?
  - Adversativity and Subject Animacy (adversative verbs with animate subjects) -> showing affectedness
  - …

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

- Logistic regression
- The model will calculate the probability that a given construction is realized with „se faire"
- The predictor variables will be normalized (log of the odds)
- Random effects (to avoid individual effects):
  - lemma of the verb
  - corpus
  - person (metadata)?
- p-value 0.05?
- 2 different models for the different modalities (spoken/written) and one combined model?
- Maybe: memory-based learning

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

- Constructions will only be included if:
  - the construction can be realized with both "se faire" and "être" passive
  - "se faire" construction has a passive (and not an exclusively causative) reading
  - the verb is a transitive verb
  - "se faire" + Inf -> exclude main verb "faire"? ("se fait"/"s'est fait" without Inf meaning "is made")
  - tense of the verb? (both temps composé and présent? → only temps composés to avoid adjectival passive? But tense could be a factor and would also lead to more data)
  - negations / complex phrases? (include but not as a factor? Complicated and probably very few constructions if any at all)
  - no modal verbs? (could be a factor but complex, potential interactions with subject responsibility, and might be very few constructions)
- …

7) How many observations will be collected or what will determine sample size?

- Sample size is probably going to be determined by the amount of “se faire” constructions that are very few in proportion.

8) Anything else you would like to pre-register?