Preregistration:

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

We will use the French open source corpus ParlaMint 2.0 and extract our sample from Sketch Engine with a semi-automatic approach (CQL and manual filtering). We have already extracted a sample of all 1,193 "se faire" constructions present in the corpus via the following CQL expression:

([lemma="se"] [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="je"] []{0,3} [lemma="me"] [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="tu"] []{0,3} [lemma="te"] [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="vous"] [tag!="Fc"]{0,3} [lemma="vous"] [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="nous"] [tag!="Fc"]{0,3} [lemma="nous"] [lemma="faire"] []{0,3} [tag="V.N.\*"])

→ We have filtered these constructions manually. After …: XXX occurences.

1. 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

1. 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 - 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”)
  + the “être” construction is not an adjectival construction (e.g. *la porte est ouverte*)
  + ~~no modal verbs? (could be a factor but complex, potential interactions with subject responsibility, and might be very few constructions) ?~~
  + Complément d’agent only with *par* and not with *de*

1. 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 (singular/plural)
    - Gender ()
  + Verb
    - TAM (tense, aspect, mode)
    - Main verb semantic group (clusters): (e.g. dynamic/stative, cognition, movement, perception…)
    - ~~Telicity of the verb (telic/atelic~~
    - Adversativity of the verb → sentiment analysis or binary annotation (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 occurrences in data)
  + ~~[Periphrase constructions (present/absent)→ probably too few in data]~~
  + Negation (present/absent) (if enough occurrences for negated constructions in the date)
* Interactions tested:
  + Animacy of subject and Complément d‘agent ?
  + Animacy and verb clusters: not all verbs can occur with animate subjects (Which ones?)
  + subject animacy and verb adversativity: Negative consequences of actions, expressed by adversative verbs, may affect animate subjects more than inanimate subjects.

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

* Mixed-effects logistic regression model
* 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
  + speaker (metadata)
* p-value 0.05
* Maximal model and backpropagation
* 1. Remove insignificant interactions
* 2. Remove insignificant variables (one by one) to find the best fitting model?

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

* If feature turns out to be irrelevant ort there’s not enough data: merge or drop?

1. How many observations will be collected or what will determine sample size?

The sample size will be determined by the amount of “se faire” passive constructions, which are very few in proportion to the “être” passive constructions. We have extracted all “se faire” passive constructions present in the corpus (see 1)) and filtered them manually to exclude false positives: After this, we ended up with xxx constructions, and this amount will be matched by the être constructions.

We will select an equal number of occurrences of the “être” passive by making a list of the lemmas and using this list to take a random sample of “être” constructions with the same lemmas. If there are not enough occurrences of these lemmas in the “être” passive constructions in the corpus, we will take a random sample of “être” passive constructions. The exact sampling process is explained in 8).

1. Anything else you would like to pre-register?

As the data are very unbalanced, we plan to take some steps to minimise bias and ensure reproducibility.

After extracting the se faire constructions we will make a list of all lemmas found in the “se faire” constructions. For the être passive constructions, we will search constructions using a CQL expression in SketchEngine. We will then filter the concordances using the list of lemmas we have previously collected from the “se faire” constructions. All filtered instances will be downloaded and stored in a .csv file. For reproducibility, we will then set a seed and use this to randomly sample the data. The seed will be determined by the lottery numbers on Wednesday, 5th March 2025 as provided on the Website <https://www.lotto.de/lotto-6aus49/lottozahlen> using the function set.seed() in R. Using this, we will take a random sample from the concordance lines that is equal in size to the manually filtered amount of se faire constructions (i.e. xxx)

Both of us will annotate all constructions manually for features listed in 4). The Inter-Annotator-Agreement will be calculated.