Preregistration:

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

We will use the French ParlaMint 2.0 corpus 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 in the corpus found via the following CQL expression:

[lemma="se"] [lemma="être"] {0,1} [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="je"] []{0,3} [lemma="me"][lemma="être"] {0,1} [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="tu"] []{0,3} [lemma="te"] [lemma="être"] {0,1}[lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="vous"] [tag!="Fc"]{0,3} [lemma="vous"] [lemma="être"] {0,1} [lemma="faire"] []{0,3} [tag="V.N.\*"] | [lemma="nous"] [tag!="Fc"]{0,3} [lemma="nous"] [lemma="être"] {0,1} [lemma="faire"] []{0,3} [tag="V.N.\*"]

We have also already manually filtered the extracted constructions. After excluding all constructions that are not “se faire” + Inf constructions and constructions that do not meet our criteria (see 3)), our sample comprises 535 occurrences.

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 spoken French (specifically: parliamentary debates)? In other words: what factors influence speakers’ choice between the “se faire” and the “être” passive construction? We operationalize this as the probability that the “se faire” passive construction is used.

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

The dependent variable is the type of passive construction (“se faire”/“être”).

We only included “se faire” constructions that have a passive (and not an exclusively causative) reading, i.e. constructions that can be paraphrased as “être” passive without becoming ungrammatical or fundamentally altering their meaning. This is our main criterion. We have defined the following additional criteria, which follow from our main criterion:

* The verb is a transitive verb (the direct object of the active sentence is the subject of the passive sentence).
* The “se faire” construction is not a fixed idiomatic expression with a figurative meaning. We especially excluded constructions classified as "idiomes lexicalisés" (lexicalized idioms) according to Raineri (2010: 163). These constructions neither carry a true causative nor passive meaning and cannot occur with a complément d’agent: *se faire entendre, se faire sentir, se faire attendre, se faire désirer* (Raineri 2010: 164).
* We exclude “être” constructions/paraphrases that are essentially adjectival constructions (e.g. *elle est connu*).

To ensure consistency and practicality, we decided to include or exclude verbs on the basis of their lemma, rather than on individual occurrences. After manually evaluating all occurrences, the lemmas that we excluded from our sample are: “entendre”, “sentir”, “ressentir”, “attendre”, “désirer”, “voir”, “connaître”, “passer (pour)”, “mousser”, “avoir”

We also exclude imperative constructions because of their rarity in the passive and because of the semantic contradiction (the passive focuses on the undergoer of an action, while the imperative demands an action). We expect that imperatives are likely to be unevenly distributed, which may lead to statistical distortions.

1. How many and which conditions will participants be assigned to?

This is a corpus study, so there will be no participants but rather texts.

The factors for the logistic regression are:

* Subject responsibility:
  + Animacy (animate/inanimate)
  + Gender (mixed/unknown/masculine/feminine)
  + Complément d’agent (absent/par/de)
* Aspectuality:
  + Tense (Past, present, future OR past/ non-past)
  + Aspect (progressive/non-progressive)
  + Modal verb construction? (no/yes)
  + Main verb semantic group (clusters): (e.g. dynamic/stative, cognition, movement, perception, telic/atelic…)
* Adversativity (non-adversative/adversative) or (positive/ neutral/ negative)
* Interactions tested:
  + Animacy of subject and Complément d‘agent: if the subject is inanimate, it more likely requires an external agent to clarify who/ what performs the action; conversely if the subject is animate, such an external agent is less needed.   
     être passives may be more common with inanimate subjects when a complement d’agent is present. Se faire passive tends to favour animate subjects without a Complément d’agent.
  + Animacy of subject and gender of the subject
  + Animacy and verb clusters: verbs of movement, psychological verbs/ emotions, communication more likely with animate subjects; states more likely with inanimate subjects?
  + Subject animacy and verb adversativity: Negative consequences of actions, expressed by adversative verbs, may affect animate subjects more than inanimate subjects.  
     Adversative verbs strengthen the preference for se faire passives whenthe subject is animate.

We recognise that the factors influencing the choice between the two passive variants are complex, and it is likely that there are meaningful interactions we have not hypothesized. Since we have not yet systematically explored the data, we think it likely that additional interactions will emerge during analysis. To reduce model complexity and improve stability, the interactions listed here are those we consider most likely to be significant to the choice between the two passive variants être and se faire.

If, during the exploratory analysis, we identify other meaningful interactions that we had not previously thought of (i.e. not listed here), we plan to further investigate them and, if appropriate, include them as factors in our model.

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

We will use a 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 using the log of the odds. To control for individual variation, we will include the verb lemma and the speaker (as indicated in the meta data) as random effects. We will set a significance threshold of p-value of 0.05.

First, we will build a maximal model that includes all the factors, interactions and random effects described in 4). For this, the glmer( ) function in R will be used.

passive\_variant ~ subj\_animacy + subj\_gender + agent\_pp + tense + aspect + mood + modal + cluster + adversativity + subj\_animacy: adversativity + subj\_animacy: gender + subj\_animacy: agent\_pp + subj\_animacy: modal + cluster: tense … + (1| lemma) + (1|speaker)

If there are any, we will then remove insignificant interactions to improve R2. Then, we will manually remove variables one by one to find the model with highest R2 (i.e. the model that explains most of the variance).

The significant factors and interactions will be visualised and analysed in terms of our research question (see 2)).

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

If a feature turns out to be irrelevant or there is insufficient data for a level of a factor:

* Animacy (animate/inanimate): If the feature turns out to be irrelevant or there is insufficient data for one of the categories, we will drop it.
* Gender (mixed/unknown/masculine/feminine):
* Complément d’agent (absent/par/de): If the feature turns out to be irrelevant, we will drop it. If there is insufficient data for the “de” and/or “par” category, we will merge these categories into a “present” category.
* Tense: (Past, present, future OR past/ non-past): If data for the “future” category is sparse, it will be collapsed with “present” into a broader **“non-past”** category.
* Aspect (progressive/non-progressive): If instances of one of the categories turn out to be rare or highly imbalanced, the factor will be removed to avoid data sparsity issues. If exploratory analysis reveals that the distinction between simple and perfect provides better explanatory power, these will be introduced as additional levels instead of the “progressive” level.
* Modal verb construction (no/yes):
* Main verb semantic group (clusters): In case of very low frequent occurrences of one cluster, it will be merged with another semantically similar cluster. If no appropriate match is found, the cluster will be dropped from the analysis to avoid model distortion.
* Adversativity of the verb (non-adversative/adversative) or (positive/ neutral/ negative): If there is insufficient data for the the “positive” or the “neutral” category, we will collapse them into a broader “non-adversative” category.

If we come across constructions which are erroneously included because they slipped past our manual filtering, these will be excluded.

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: After this, we ended up with 535 constructions. We will select an equal number of occurrences of the “être” passive by making a list of the lemmas of the “se faire” passive occurrences and use this list to collect 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, 12th March 2025 at 19:25 pm 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. 535).

Both of us will annotate all constructions manually for features listed in 4). The Inter-Annotator-Agreement will be calculated for the verb cluster and adversativity annotations.