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 1,254 "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?

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.

Raineri, S. (2010). Analyse contrastive français-anglais du passif dans une perspective constructionnelle :Senset fonction deBEVen,ETREVé,GETVenetSEFAIREVer [PhD thesis, Université de la Sorbonne nouvelle - Paris III]. <https://theses.hal.science/tel-00841659>

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)
  + Modal verb construction (yes/no)
* Aspectuality:
  + TAM (progressive/non-progressive)
  + Main verb semantic group/cluster (hitting-touching/communication/neutral)
* Adversativity (adversative/non-adversative)
* Interactions tested:
  + Animacy of subject and gender of the subject: se faire could be more frequent for animate masculine subjects than for animate feminine subjects. For inanimate subjects, gender could have less influence.
  + Animacy and verb clusters: The choice of passive construction could be influenced not only by animacy or verb cluster alone, but by the combination of these two factors: e.g. animacy could have a stronger effect if the verb is from the hitting-touching category or the neutral category (“se faire” could be strongly favoured with animate subjects compared to inanimate subjects), while this effect could be weaker with communicative verbs for which “être” passive could be favoured regardless of animacy.
  + Subject animacy and verb adversativity: negative consequences of actions (expressed by adversative verbs) could be perceived differently depending on the animacy of the subject. Adversative verbs could affect animate subjects and favour se faire more strongly.
  + **Verb cluster and adversativity:**Adversativity could reinforce se faire more in communication and neutral verbs than in hitting-touching verbs: Communication verbs tend to favour être and neutral verbs have no strong preference, but adversative constructions may shift both toward se faire. Since hitting-touching verbs are inherently highly adversative, adversativity is unlikely to strongly affect their passive preference.

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

We will fit a mixed-effects binomial logistic regression model in R following Chapter 12 (pp. 253-276) in Levshina (2015). The model will calculate the log of the odds that a given construction is realized with „se faire“. 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).

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

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

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

We plan to check the model for overfitting using bootstrap validation, following Levshina (2015: pp. 166-167; 274-275).

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.195>

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): If the feature turns out to be irrelevant, we will drop it. If there is insufficient data for one of the categories, we will drop it.
* 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.
* TAM (progressive/non-progressive): if instances of one of the categories turn out to be rare or highly imbalanced, the factor will be removed.
* Modal verb construction (yes/no): if the feature turns out to be irrelevant or there is insufficient data for one of the categories, we will drop it.
* Main verb semantic group/cluster (hitting-touching/communication/neutral): if the feature turns out to be irrelevant, we will drop it. In case of very few occurrences in the communication and/or neutral cluster, these will be merged.
* Adversativity of the verb (adversative/non-adversative): if the feature turns out to be irrelevant or there is insufficient data for one of the categories, we will drop it.

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 18:25 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). The constructions will be manually annotated for the features listed in 4).