Yanis da Cunha
Université Paris Cité & Laboratoire de Linguistique Formelle (France)

Anne Abeillé
Université Paris Cité & Laboratoire de Linguistique Formelle (France)

L'alternance actif/passif en français parlé : un modèle statistique

1. INTRODUCTION [1]

1.1. Le passif à l'oral

La fréquence de la construction passive varie selon les genres de texte, le passif étant plus fréquent à l'écrit qu'à l'oral, aussi bien en anglais (Roland, Dick & Elman 2007) qu'en français (Poiret & Liu 2020). À cette première différence s'ajoute la question des facteurs qui conditionnent l'utilisation d'un passif. En anglais, des facteurs communs entre écrit et oral ont été trouvés, suggérant une certaine unité de la construction à travers les genres textuels (Bresnan, Dingare & Manning 2001 ; Estival 1985 ; Hundt, Röthlisberger, Seoane 2018 ; Weiner & Labov 1983). Pour le français écrit, des facteurs similaires ont été mis en évidence par Y. da Cunha et A. Abeillé (2020). Toutefois, les facteurs conduisant à l'utilisation du passif en français parlé n'ont pas encore fait l'objet d'une description quantitative.

Plusieurs études du passif en français oral ont été menées sur corpus. C. Blanche-Benveniste (2000), en comparant le corpus oral du GARS (Groupe Aixois de Recherche en Syntaxe) et des textes journalistiques, met en évidence le rôle de la classe sémantique des verbes : les verbes psychologiques favorisent l'utilisation du passif (1). L'étude de B. Hamma, A. Tardif et F. Badin (2017), sur le corpus ESLO (Enquêtes Sociolinguistiques à Orléans), permet quant à elle de faire l'hypothèse d'un effet de structure informationnelle, dans la mesure où le

1. Nous remercions pour leurs commentaires les relecteurs de la revue *Langue française*, ainsi que les participants du colloque sur le français parlé à Nancy en octobre 2021, et ceux des colloques en ligne LSRL 2021 et NWAV 2021. Ce travail a bénéficié du soutien du LabEx EFL (ANR-10-LABX-0083).

sujet des passifs serait majoritairement pronominal (2), tandis que le complément d'agent correspond à une information nouvelle (Hamma 2015). Enfin, R. Druetta (2020), explorant le corpus Oral de Français de Suisse Romande (OFROM), montre que le passif court, malgré l'omission de son argument agentif, permet souvent son inférence en contexte (3) :

(1) J'ai été surprise du comportement des Suisses allemands (GARS – Genev 99,1)

(2) On a été soutenu par un monsieur aussi c'est monsieur NPERS (ESLO)

(3) Deux Valaisannes$_i$ [...] avaient tout organisé [...] c'était bien organisé$_i$ (OFROM)

Ces études permettent donc d'envisager l'utilisation du passif comme un phénomène essentiellement multifactoriel, mais elles présentent plusieurs limites méthodologiques. Notamment, elles ne comparent pas systématiquement construction active et passive ni n'emploient de méthode quantitative, rendant difficile la caractérisation de l'alternance de construction et des facteurs conduisant à l'utilisation de la variante passive. C'est donc par une approche quantitative, multifactorielle et probabiliste que nous proposons d'étudier l'alternance actif/passif en français oral.

1.2. Une approche probabiliste

La grammaire offre régulièrement aux locuteurs des façons alternatives d'exprimer un même message (« alternate ways of saying ‹the same› thing », Labov, 1972 : 188), comme c'est le cas dans l'alternance entre actif et passif [2]. L'approche probabiliste de la grammaire considère que ces variantes alternantes ne sont pas distribuées de façon catégorielle mais selon des préférences syntaxiques gradientes, probabilistes, qui peuvent varier selon les locuteurs, les variétés de langue, les genres de textes, entre autres. Ces préférences syntaxiques s'expriment par des contraintes préférentielles, qui orientent le choix pour une variante syntaxique donnée selon des facteurs multiples et hétérogènes (Thuilier 2012b). L'idée que la grammaire comporte ainsi une composante probabiliste définit le champ de la syntaxe quantitative (Bîlbîie, Faghiri & Thuilier 2021 ; Bresnan & Ford 2010 ; Szmrecsanyi *et al.* 2017), dont les méthodes consistent en l'usage d'outils statistiques pour l'analyse de données de corpus ou d'expérience.

Pour l'alternance entre actif et passif en français écrit journalistique (FRENCH TREEBANK, Abeillé, Clément & Liégeois 2019), Y. da Cunha et A. Abeillé (2020) montrent ainsi que le choix de construction est soumis à des contraintes variées :

2. Nous ne tenons pas compte ici de la construction passive en *se* (*Ce roman se vend bien*), décrite notamment par Desclés & Guentchéva (1993) et Zribi-Hertz (1982, 2008).

– contrainte de longueur croissante des constituants : l'actif et le passif sont utilisés de façon à placer un sujet court avant les compléments plus longs (Wasow 2002) ;
– contrainte de codage harmonique des arguments : les sujets tendent à être définis, animés, pronominaux, et le choix de construction permet de maintenir cette préférence de codage des arguments (Aissen 1999, 2003) ;
– contrainte d'amorçage : l'utilisation antérieure d'un passif dans le discours favorise sa réutilisation ultérieure (Bock 1986) ;
– contrainte de sémantique lexicale : les verbes de communication tendent à s'utiliser à l'actif et les verbes de contact (frapper, détruire, assassiner) au passif.

Ces différentes contraintes influencent simultanément le choix de construction, rendant compte du caractère gradient et multifactoriel de l'alternance actif/passif. Pour l'essentiel, nous testons les mêmes facteurs en français parlé, de façon à voir dans quelle mesure le changement de médium entraîne un changement, ou non, dans les contraintes préférentielles pesant sur l'alternance. Grâce à la méthode de la modélisation statistique des données de corpus, nous pourrons quantifier l'effet des facteurs et ainsi prédire l'utilisation des différentes variantes.

2. MÉTHODOLOGIE

2.1. Corpus choisis

Nous travaillons sur trois corpus oraux tirés du projet ORFÉO (Debaisieux & Benzitoun 2020), à savoir le *Corpus de Français Parlé Parisien* (CFPP2000, Branca-Rosoff *et al.* 2012), le *Corpus de Référence de Français Parlé* (CRFP, Delic 2004) et le *Corpus Oral de Langues Romanes* français (C-ORAL-ROM, Cresti & Moneglia 2005). Ces corpus présentent des productions orales principalement de conversation spontanée, de niveaux de formalité variés, et d'origines géographiques diverses en France (Paris pour le CFPP2000, 37 villes de France métropolitaine pour le CRFP). L'empan temporel des corpus va de 1994 à 2012. Ces corpus ont été annotés en dépendances (Kahane *et al.* 2017), ce qui permet de chercher aisément les structures passives et actives (Fig. 1).

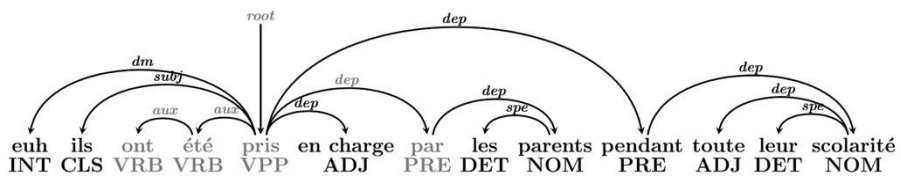

Figure 1 : Arbre de dépendance pour la phrase *euh ils ont été pris en charge par les parents pendant toute leur scolarité* (Pierre-Marie Simo, CFPP2000) ; le verbe ÊTRE a la fonction AUX et le complément d'agent DEP

2.2. Échantillonnage

Suivant la méthode de Y. da Cunha et A. Abeillé (2020), nous optons pour un échantillon équilibré contenant autant de passifs (200 énoncés) que d'actifs (200 énoncés), pour un total de 400 énoncés. Le groupe des passifs comporte une moitié de passifs avec complément d'agent en *par* (4a), que nous appellerons *passifs longs*, et une moitié de passifs sans complément d'agent (4b), que nous appellerons *passifs courts*. En correspondance à l'actif, nous sélectionnons une moitié d'actifs à sujet SN (5a) et une moitié d'actifs à sujet clitique (5b) :

(4) a. [J']$_{ARG2}$ **avais été enthousiasmé** [par le premier d'ouverture]$_{ARG1}$ (REN, fmedsp02, C-ORAL-ROM)
b. Saint-Antoine c'est là où [ma mère]$_{ARG2}$ **a été opérée** (CFPP2000)

(5) a. Et [ma mère]$_{ARG1}$ m'**a raconté** [quelque chose d'extraordinaire]$_{ARG2}$ (L1, PRI-POI-2, CRFP)
b. Donc il m'a expliqué [il]$_{ARG1}$ m'**a donné** [des petites ficelles]$_{ARG2}$ (L1, PRI-PSO-2, CRFP)

L'échantillon extrait ne contient que des verbes transitifs passivables à arguments nominaux ou pronominaux et fléchis à un temps composé (afin d'éviter l'ambiguïté entre passif adjectival et verbal aux temps non composés). Sa composition est détaillée dans le Tableau 1.

Tableau 1 : Nombre d'énoncés de l'échantillon étudié par construction et par corpus

	CFPP	CRFP	C-ORAL-ROM	*Total*
Actif à sujet clitique	30	49	21	*100*
Actif à sujet SN	30	49	21	*100*
Passif court	30	49	21	*100*
Passif long	30	49	21	*100*
Total	*120*	*196*	*84*	*400*

2.3. Facteurs annotés

Nous décrivons l'alternance actif/passif en désignant le sujet actif et le complément d'agent du passif par ‹argument 1› (Arg1) et l'objet actif et le sujet passif par ‹argument 2›. Nous annotons alors les 11 facteurs suivants :

– le genre grammatical, le nombre, la personne [3], la définitude (défini, indéfini) et la catégorie syntaxique (nom, pronom) des arguments ;

3. En cas de *mismatch* entre personne grammaticale et référentielle, comme dans le cas de *on*, nous faisons le choix de privilégier la personne référentielle. L'exemple suivant est donc annoté avec un ‹argument 2› de 1re personne :
On a été remarqué par un producteur [...] (L1, PRI-PSO-4, CRFP)

– l'animéité des arguments, en utilisant les catégories de A. Zaenen *et al.* (2004) ;
– la longueur des arguments en nombre de mots (Thuilier 2012b) ; nous définissons alors une variable de différence de longueur, qui utilise les valeurs de longueur des arguments passées au logarithme, afin de tenir compte de la distribution exponentielle du nombre de mots : Différence de Longueur = *log(longueur Arg1) – log(longueur Arg2)* ;
– le lemme et la classe sémantique du verbe, à l'aide du dictionnaire *Les Verbes Français – LVF* (Dubois & Dubois-Charlier 1997) ;
– le type syntaxique de la phrase (principale, subordonnée) ;
– la présence d'un passif dans les trois phrases qui précèdent (dans leurs corpus oraux en anglais, Estival (1985) et Weiner & Labov (1983) regardaient jusqu'à cinq phrases précédentes).

2.4. Analyse statistique

Nous modélisons les données à l'aide de la régression logistique à effets mixtes (package LME4 sur R, Bates *et al.* 2014). Ce type de régression permet de quantifier la relation entre un ensemble de variables prédictives et une variable binomiale à prédire. Le modèle ainsi obtenu permet de prendre en compte la direction, la force et la significativité statistique d'effets multiples sur la variable à prédire. Dans notre cas, la variable à prédire est la variable VOIX, qui a deux valeurs : actif (VOIX=0) ou passif (VOIX=1). Les variables prédictives correspondent aux facteurs annotés (les variables catégorielles ont été normalisées). Les variables LEMME et CORPUS sont utilisées comme variables aléatoires. Ces variables représentent des sous-groupes de données qui se forment de façon aléatoire par échantillonnage dans la population. Les inclure comme variables aléatoires dans le modèle permet de calculer des effets aléatoires, qui correspondent à un coefficient d'*intercept* spécifique à chaque lemme, chaque corpus. De cette façon, on rend compte du bruit aléatoire induit par les sous-groupes de données, par exemple de la préférence lexicale individuelle d'un lemme pour l'actif ou le passif.

Une fois qu'un modèle a appris une relation entre variables à partir d'un échantillon, celui-ci est capable de généraliser. Le modèle peut ainsi être utilisé de façon prédictive, et il fournit alors une probabilité de passif P pour tout énoncé annoté. Si $P > 50\ \%$, le modèle prédit un passif et si $P \leq 50\ \%$, le modèle prédit un actif. Pour évaluer la capacité prédictive d'un modèle, nous utilisons une procédure de validation croisée répétée 100 fois (package CARET sur R, Kuhn 2008). La procédure permet d'obtenir une mesure d'exactitude qui reflète la capacité du modèle à prédire si un énoncé correspond à un actif ou à un passif sur des données inconnues. Cette mesure doit être comparée à la *baseline* du modèle, qui correspond à l'exactitude avec l'*intercept* comme seul prédicteur [4].

4. Le corpus annoté et le script permettant la reproduction des analyses statistiques sont disponibles sur une archive ouverte accessible au lien suivant : https://osf.io/udg36/.

Tableau 2 : Modèle de régression logistique du passif long (300 points de données)

	Coefficient	Écart-type	z-value	p-value
Intercept	-1.82	0.33	-5.51	< 0.001
Différence de longueur	1.88	0.38	4.99	< 0.001
Argument1 pronominal	-0.92	0.38	-2.42	0.01548
Argument1 défini	-0.79	0.26	-3.06	0.00219
Argument1 humain	-0.75	0.22	-3.48	0.00050
Argument2 défini	0.72	0.28	2.57	0.01012
Verbe de type *frapper*	0.57	0.25	2.30	0.02129
Passif antérieur	0.46	0.22	2.05	0.04028
Exactitude = 85 % (Baseline = 67 %)				

Par la prise en compte de ces différents facteurs, le modèle peut assigner une valeur de probabilité de passif P aux phrases, rendant compte du caractère gradient et probabiliste de l'alternance actif/passif. Les phrases en (6) sont ainsi correctement classifiées comme des passifs (P > 50 %), tandis que les phrases en (7) sont correctement classifiées comme des actifs (P $\leq$ 50 %) :

(6) a. [P = 95 %] est-ce que il vous semble que vous avez été touché par la crise économique ou pas spécialement ? (CFPP2000)
b. [P = 97 %] ce geste aurait été interprété d'une façon totalement positive par la plupart des formations politiques (L12, PUB-BAY-1, CRFP)

(7) a. [P = 1 %] les commerçants avaient fait une petite animation (CFPP2000)
b. [P = 2 %] si vous m'aviez donné la possibilité de préparer je pense que j'aurais été plus claire (L2, PRI-PRI-2, CRFP)

En (6a), l'utilisation du passif a une probabilité de 95 %. Plusieurs facteurs contribuent à cela : l'argument1 est non humain et plus long que l'argument2, l'argument2 est défini, et le verbe utilisé appartient à ceux de la classe des verbes de type *frapper* (classe F2b2 dans le *LVF*). En (7a), l'utilisation du passif est peu probable (P = 1 %) et l'actif est donc prédit à 99 %. On trouve cette fois un argument1 humain, plus court que l'argument2 et l'argument2 est indéfini, facteurs favorisant l'actif.

En prédiction sur ses données d'apprentissage, le modèle a une exactitude de 90 %, ce qui est le signe d'un léger surajustement. La matrice de confusion suivante montre comment le modèle classifie les données :

Tableau 3 : Matrice de confusion du modèle du passif long

	Actif observé	Passif observé	Exactitude
Actif prédit	188	19	91 %
Passif prédit	12	81	87 %
Exactitude générale			90 %

Le modèle classifie donc incorrectement 31 phrases (10 % des données).

(8) a. [P = 36 %] j'ai toujours **été rebuté** par ça (L1, PRI-BEL-1, CRFP)
b. [P = 93 %] ces fameuses lettres **avaient blessé** son amour-propre (LES, fnatla02, C-ORAL-ROM)

En (8a), le modèle aurait trouvé plus probable l'actif (P > 50 %) : *ça m'a toujours rebuté*, tandis qu'en (8b), il prédisait un passif (P < 50 %) : *son amour-propre avait été blessé par ces fameuses lettres*. La phrase (8a) met en jeu deux arguments pronominaux, et de même longueur, ce qui écarte cette phrase du schéma fréquent du passif ‹[Pronom défini] *a été* V *par* [Nom indéfini]›. La phrase (8b) met, quant à elle, en jeu deux arguments nominaux inanimés et définis, ce qui s'écarte du schéma d'actif le plus fréquent ‹[Pronom défini animé] *a* V [Nom indéfini inanimé]›. Par sa classification incorrecte, le modèle met donc en exergue les cas plus marginaux et montre que les contraintes qui pèsent sur l'alternance sont gradientes et probabilistes, mais aussi stochastiques. Cela peut aussi indiquer que d'autres facteurs seraient à prendre en compte.

Nous proposons d'étudier chaque type d'effet à l'aide de statistiques descriptives, avant de présenter un modèle du passif court.

2.5. La contrainte de longueur croissante des constituants

Un premier effet significatif que nous mettons en évidence est celui de la longueur : plus l'argument1 est long par rapport à l'argument2, plus le passif est fréquent (E = 1.88 ; SE = 0.38 ; $p < 0.001$) [5]. La Figure 2 permet de voir cette relation. Lorsque l'argument1 est plus court que l'argument2 (différence de longueur < 0), le passif est peu fréquent et l'on utilise donc principalement l'actif, c'est-à-dire un ordre Arg1–Arg2. Lorsque que l'argument1 est plus long que l'argument2, (différence de longueur > 0), on voit le passif émerger, c'est-à-dire un ordre Arg2–Arg1. Autrement dit, de manière générale, on choisit la construction qui permet à l'argument le plus court de précéder l'argument le plus long. Cela correspond à la contrainte générale de longueur croissante des constituants (Eitelmann 2016 ; Wasow 2002). Comme le suggère la Figure 2, cette relation entre longueur et fréquence peut être modélisée par une courbe de régression logistique (en bleu). Ce même type de relation a été mis en évidence pour l'ordre des compléments post-verbaux en français (Thuilier 2012b) et pour l'alternance actif/passif en français écrit (da Cunha & Abeillé 2020), suggérant que de façon générale la

5. E : estimate (valeur estimée du coefficient) ; SE : standard-error (erreur-type).

probabilité d'un ordre X–Y entre deux constituants peut être exprimée comme une réponse logistique à la différence *log(longueur de Y) – log(longueur de X)*. Dans notre cas, la probabilité de l'ordre Arg2–Arg1 (le passif) est fonction de la différence *log(longueur de Arg1) – log(longueur de Arg2)*. Notons, par ailleurs, que lorsque les deux arguments ont la même longueur (différence de longueur = 0), on observe peu de passifs (seulement 17 %), et l'actif apparaît donc comme la construction majoritaire par défaut sans influence de ce facteur.

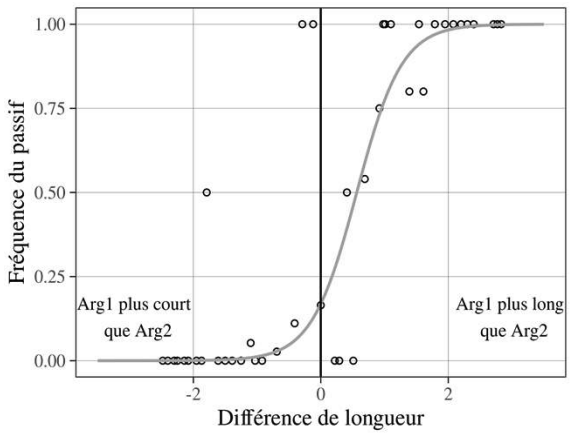

Figure 2 : Courbe logistique exprimant la fréquence du passif en fonction de la différence de longueur des arguments *log(longueur de Arg1) – log(longueur de Arg2)*

2.6. La contrainte de codage harmonique des arguments

D'autres propriétés des arguments que la longueur jouent un rôle dans nos données. Le modèle met notamment en lumière des effets de pronominalité (Arg1 pronominal : E = -0.92 ; SE = 0.38 ; $p < 0.05$), d'humanitude (Arg1 humain : E = -0.75 ; SE = 0.22 ; $p < 0.001$) et de définitude (Arg1 défini : E = -0.79 ; SE = 0.26 ; $p < 0.01$ et Arg2 défini : E = 0.72 ; SE = 0.28 ; $p < 0.05$). Ces propriétés peuvent être représentées sous la forme d'échelle de proéminence :

(9) Échelle de proéminence
Pronom > Nom
Défini > Indéfini
Humain > Non-humain

Les valeurs proéminentes de ces échelles (défini, humain, pronom) représentent les référents qui sont plus faciles à récupérer en mémoire et qui, de ce fait, tendent à être codés plus tôt dans la phrase ou avec une fonction plus proéminente comme la fonction sujet. Nous formulons ce phénomène comme une contrainte de codage harmonique des arguments (*easy first principle* (MacDonald 2013) ; *harmonic alignment* (Aissen 1999) ou encore *role-reference*

association (Haspelmath 2020)). Les effets significatifs de notre modèle vont dans le sens de cette tendance : les arguments1 pronominaux, humains et définis augmentent significativement la fréquence de l'actif, et les arguments2 définis celle du passif, c'est-à-dire que ces arguments apparaissent plus fréquemment dans la construction où ils sont sujets. Ce résultat est complété par ceux de la Figure 3, qui représente la fréquence du codage sujet selon les valeurs de définitude, d'humanitude et de pronominalité. On voit que pour l'argument1 comme l'argument2, avoir une valeur proéminente (défini, humain, pronom) correspond à une augmentation de la fréquence du codage sujet par rapport à la valeur non proéminente (indéfini, non humain, non pronom).

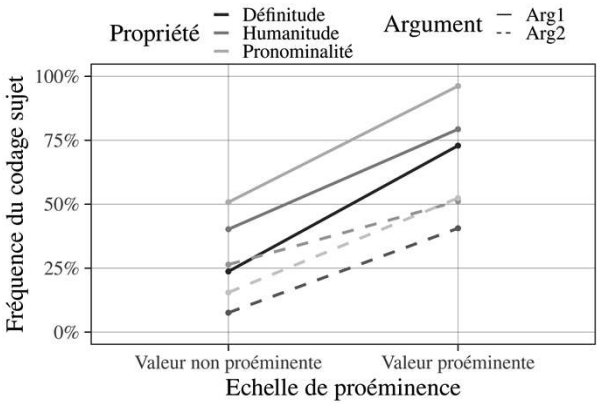

Figure 3 : Effet des échelles de proéminence sur la fréquence du codage sujet ; les valeurs prééminentes des échelles (défini, humain, pronom) augmentent la fréquence du codage sujet

En somme, nous montrons donc que le choix de construction dans l'alternance entre actif et passif vise à favoriser les codages harmoniques des arguments, de sorte que les sujets soient définis, humains et pronominaux et les compléments indéfinis, non humains et nominaux. Cela explique que (10a) ait une probabilité de passif plus importante que (10b) :

(10) a. [P = 82 %] Le tombeau **a été ouvert** par les autorités religieuses (L1, PRI-POI-2, CRFP)
b. [P = 49 %] Des crédits **ont été alloués** par l'Europe par euh la municipalité et l'État (L1, PRI-TRO-2, CRFP)

Le même effet d'humanitude a par ailleurs été observé en expérience en français dans une tâche de rappel de phrases, avec le résultat qu'un argument2 humain favorise le passif, c'est-à-dire le codage sujet (Thuilier *et al.* 2021).

2.7. L'effet des classes sémantiques de verbe

Notre modèle montre également la significativité d'un effet de sémantique lexicale : les verbes de type *frapper* augmentent la fréquence du passif (E = 0.57 ;

SE = 0.25 ; p < 0.05). Ainsi, 85 % des verbes de type *frapper* (23 énoncés) sont utilisés au passif comme en (11) :

(11) a. bon Paris a été très peu bombardé hein (Yvette Audin, CFPP2000)
b. [P = 96 %] il a été ravagé par les flammes (Raphael Lariviere, CFPP2000)
c. [P = 70 %] toute leur culture leur religion a été détruite par les Espagnols (L1, PRI-LEM-1, CRFP)

L'effet n'étant pas catégorique, on trouve également des énoncés à l'actif avec des verbes de cette classe sémantique (15 % des verbes de type *frapper*, soit 4 énoncés) :

(12) [P = 43 %] les Anglais ont menacé la France d'une guerre en Europe (FRE, ffammn17, C-ORAL-ROM)

Cet effet est comparable à ce qui a été trouvé en anglais journalistique avec les verbes dénotant des actions violentes, qui sont également plus fréquents au passif (Henley, Millier & Baezley 1995). Cette préférence des verbes de type *frapper* pour la construction passive peut s'expliquer, d'une part, par le caractère topical du patient affecté et, d'autre part, par la possibilité d'effacer un agent inconnu ou peu saillant, comme en (11a).

La Figure 4 donne à voir la répartition des constructions actives et passives pour chaque classe sémantique de verbe dont la fréquence représente plus de 5 % des données (20 occurrences).

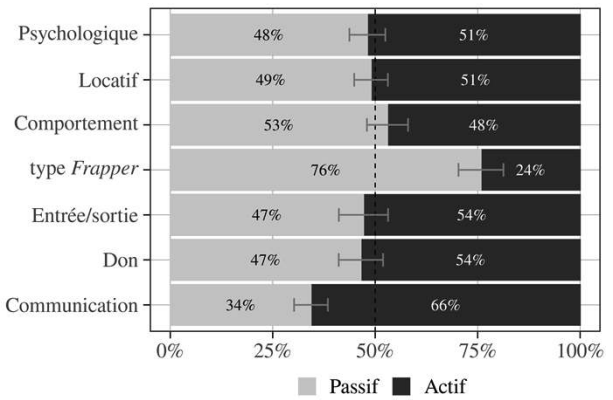

Figure 4 : Fréquence des constructions actives et passives par classe sémantique

On constate que certaines classes sémantiques n'ont pas d'effet sur la fréquence du passif (50 % de chaque construction environ) : les verbes de saisie, de réalisation, psychologiques, de comportement, d'entrée/sortie et de don. Les verbes de type *frapper* sont plus fréquents au passif (85 %) comme décrit précédemment. On constate aussi que les verbes locatifs et de communication semblent plus fréquents à l'actif. Dans un modèle de régression logistique à un

seul facteur (modèle univarié), l'écart des verbes locatifs n'apparaît pas significatif (E = -0.11 ; SE = 0.15 ; z = -0.71 ; p = 0.48) ; en revanche, celui des verbes de communication est significatif (E = -0.5 ; SE = 0.16 ; z = -3.16 ; p = 0.0016). Les verbes de communication sont donc significativement plus fréquents à l'actif. Une hypothèse pour l'expliquer est celle de la disposition verbale (Stallings, MacDonald & O'Seaghdha 1998). Cette hypothèse avance que les verbes ont des biais pour certaines constructions, qui sont stockés par les locuteurs dans leur lexique mental. Ces biais influencent alors l'emploi des verbes. En l'occurrence, les verbes de communication ont la particularité de se construire avec des complétives objets et ils auraient donc un biais pour se construire avec un objet en général (les complétives étant rarement sujet). Or, comme l'actif permet au verbe d'avoir un objet, les verbes de communication tendent à maintenir cette construction active avec objet, et non la construction passive. Une hypothèse similaire avait été proposée par J. Thuilier (2012a) concernant les verbes de communication et les compléments post-verbaux. Dans le cas du passif, cette explication resterait à corroborer par une analyse plus approfondie, par exemple en réalisant un échantillon par classe sémantique de verbe. Remarquons enfin que les verbes de communication mettent en jeu un argument1 animé ou humain, qui est plus fréquemment codé comme sujet à l'actif (cf. § 3.2).

2.8. La contrainte d'amorçage

Pour terminer, notre modèle inclut un effet d'amorçage (Bock 1986) : il est plus probable d'utiliser un passif si un autre passif est présent dans le discours antérieur (E = 0.46 ; SE = 0.22 ; $p < 0.05$). Le Tableau 4 présente nos résultats quant à ce facteur. On voit ainsi que si une amorce passive est présente dans une des trois phrases qui précèdent, la fréquence du passif est de 70 % contre 30 % pour l'actif. En l'absence d'amorce dans le contexte, c'est l'actif qui est plus fréquent (57 %), se plaçant à nouveau comme construction par défaut. Ce résultat est cohérent avec les études en anglais oral (Estival 1985 ; Weiner & Labov 1983) et en français écrit (da Cunha & Abeillé 2020).

Tableau 4 : Effet de l'amorçage sur l'alternance actif/passif

	Amorce passive		Pas d'amorce	
Actif	31	30 %	169	57 %
Passif	74	70 %	126	43 %
Total	*105*	*100 %*	*295*	*100 %*

Notons que l'effet relevé ici ne peut toutefois pas recevoir d'explication unique. Plusieurs facteurs peuvent l'expliquer : reprise du même verbe ou de la préposition (par amorçage lexical, répétition), reprise du même sujet (cohésion textuelle) ou reprise de la même construction (amorçage syntaxique). L'effet d'amorçage passif peut donc s'expliquer aussi bien en termes cognitifs (réutilisation facilitée d'éléments déjà mentionnés) ou discursifs (construction du

discours par des effets de reprise). À titre d'exemple, on trouve des cas où une portion de discours entière consiste en des répétitions lexicales et syntaxiques (13). L'effet de l'amorçage nécessiterait donc, sans doute, d'être étudié plus avant, par des protocoles expérimentaux ou une annotation en corpus plus fine par exemple :

(13) Puisque c'est une civilisation qui parle maintenant espagnol puisqu'ils **ont été envahis** par les espagnols toute leur culture leur religion **a été détruite par** les espagnols en quinze cent et quelques [...] et tous ces sites **ont été découverts** fin du siècle dernier début du siècle début du vingtième siècle parce qu'ils **étaient complètement envahis par** la la forêt amazonienne complètement ils **ont vraiment été découverts** par hasard hein (L1, PRI-LEM-1, CRFP)

3. UN MODÈLE POUR LE PASSIF COURT

Le Tableau 2 présentait un modèle du passif long, qu'il est plus facile de manipuler car le passif long ne présente pas de valeur vide pour l'argument1. À l'inverse, l'argument1 omis du passif court n'a pas fait l'objet d'une annotation complète pour toutes nos variables. Notamment, la longueur, la catégorie syntaxique et la définitude n'ont pas été annotées, tandis que la personne, le genre, l'animéité et le nombre ne l'ont été que lorsque l'argument était inférable dans le contexte. Comme R. Druetta (2020) le montrait dans le corpus OFROM, l'argument1 omis du passif court est souvent présent dans le discours. On peut qualifier cette récupération d'*endophorique* (49 % des cas). En (14), ce sont les indiens qui auraient mal reçu le Pape :

(14) L2 : Les Indiens$_i$ étaient absolument eux fascinés de voir que le Pape vienne
L1 : Oui j'avais entendu dire qu'il **avait été mal reçu**$_i$ au contraire (PRI-VAL-2, CFRP)

L'inférence peut également être *exophorique* (16 % des cas), passant par les connaissances extralinguistiques, lorsque l'agent est évident ou spécialisé dans la réalisation d'un procès (Hamma, Tardif & Badin 2017). En (15), on comprend ainsi que ce sont les médecins ou chirurgiens qui sont agents :

(15) Saint-Antoine c'est là où ma mère **a été opérée** (CFPP2000)

On trouve cependant, dans 31 % des cas, un argument1 omis avec une interprétation vague, indéfinie (type *on*, *quelqu'un*), sans que l'on puisse savoir quel agent est impliqué (16). Toutefois, la situation décrite implique quand même un agent qui n'est donc pas supprimé de l'interprétation :

(16) la cave de Valserre la cave de Monêtier Allemond et la cave de Remollon [...] toutes les trois **ont été construites** dans les années cinquante (L1, PRO-GAP-1, CRFP)

Enfin, plus rarement (4 % des cas), le passif court crée une interprétation médiopassive (Desclés & Guentchéva 1993), où l'agent est complètement supprimé de l'interprétation. En (17), les situations décrites n'impliquent aucun agent :

(17) a. Vous **avez été baigné** aussi quand même dans un milieu plurilingue (André Morange, CFPP2000)
b. Certains parents **ont été confrontés** parfois [à] des difficultés euh légères (L1, PUB-PCR-1, CRFP)

Ainsi, bien que le passif court omette l'argument1, celui-ci reste interprété dans 96 % des cas. Le modèle que nous proposons pour le passif court se trouve dans le Tableau 5.

Tableau 5 : Modèle de régression logistique du passif court (296 points de données)

	Coefficient	Écart-type	z-value	p-value
Intercept	-0.17	0.21	-0.81	0.41802
Longueur de l'argument2	-1.51	0.35	-4.32	0.00002
Argument2 défini	0.74	0.25	2.91	0.00359
Argument1 de 1re/2e pers.	-0.54	0.16	-3.36	0.00078
Passif antérieur	0.63	0.15	4.07	0.00005
Verbe de type *frapper*	0.36	0.18	1.98	0.04811
Exactitude = 73 % (Baseline = 68 %)				

Plusieurs remarques peuvent être faites concernant ce modèle. Tout d'abord, on retrouve essentiellement des facteurs communs avec le modèle du passif long, à l'exception de ceux qui impliquent l'annotation de l'argument1 omis (définitude, longueur, pronominalité). On trouve ainsi à nouveau des effets significatifs de longueur (si l'argument2 est long, l'actif est plus fréquent), de définitude (un argument2 défini augmente la probabilité de passif), d'amorçage et de classe sémantique du verbe (avoir une amorce passive ou un verbe de type *frapper* augmente la probabilité du passif). On trouve également un nouvel effet, celui de la personne : l'actif est significativement plus fréquent avec des arguments1 de 1re/2e personne. Ce phénomène correspond à la contrainte de codage harmonique respectant l'échelle de proéminence suivante :

(18) Personne de discours (1re/2e) > 3e personne

Cette échelle implique que les arguments de 1re/2e personne soient préférablement codés comme sujets (Bresnan, Dingare & Manning 2001). Cela correspond à l'effet observé dans nos données pour l'argument1 de 1re/2e personne, qui est plus fréquent à l'actif, c'est-à-dire comme sujet. Ce modèle permet donc de voir des similitudes entre passifs longs et courts. Remarquons toutefois que le modèle du passif court a de moins bonnes performances prédictives. Il ne permet en

effet de gagner que 5 points d'exactitude par rapport à la *baseline*, alors que le modèle du passif long permettait d'en gagner 18.

4. FRANÇAIS ÉCRIT ET ORAL

Comme indiqué en introduction, une première façon de comparer le passif à travers les médiums et les genres est de s'intéresser à la fréquence de la construction. Le Tableau 6 présente les résultats pour le FRENCH TREEBANK (FTB, Abeillé, Clément & Liégeois 2019) et les corpus oraux de notre étude (CO), auxquels nous ajoutons, à titre de comparaison, les résultats de R. Poiret et H. Liu (2020) issus des corpus *Surface Universal Dependencies* (SUD) [6].

Tableau 6 : Fréquence du passif dans plusieurs corpus oraux et écrits

	Corpus écrits		Corpus oraux	
	FTB	SUD Sequoia	CO	SUD Spoken-French
Nombre de passifs	3 763	343	4 210	145
Ratio Passif/Verbe	7.2 %	11.8 %	2.4 %	4.1 %
Ratio Passif/Token	6.6 ‰	11.8 ‰	3.7 ‰	5 ‰

On voit ainsi que le passif apparaît plus fréquemment à l'écrit, de façon cohérente à travers les corpus. Toutefois, nous nous garderons de conclure que cette différence de fréquence soit le reflet d'une différence grammaticale. Elle ne pourrait en effet être due qu'à un changement d'habitat textuel (Szmrecsanyi 2016), c'est-à-dire que la fréquence des contextes favorisant le passif pourrait changer entre écrit et oral sans pour autant que ces contextes soient différents selon le médium. Une autre façon de traiter la variation est ainsi de s'intéresser à des changements de grammaire probabiliste, ce qui correspond à des différences dans les contraintes qui conditionnent l'usage des constructions.

Pour ce faire, nous comparons l'alternance actif/passif dans le FTB (da Cunha & Abeillé 2020) et dans nos corpus oraux à l'aide de régressions logistiques qui calculent les interactions entre facteurs, permettant de voir si l'effet d'un facteur dépend d'un autre facteur (ici, le médium). Le Tableau 7 présente les différences significatives que nous avons observées quant à la fréquence du codage en sujet.

6. Le corpus écrit SUD_Sequoia contient de l'écrit journalistique, technique et de WIKIPÉDIA (Candito & Seddah 2012). Le corpus oral SUD_Spoken-French est une conversion du corpus RHAPSODIE (Lacheret *et al.* 2014).

Tableau 7 : Différences significatives de poids des facteurs entre FTB et CO

Facteur	Fréquence du codage sujet		Interactions avec le médium		
	CO	FTB	coef.	Erreur-type	p-value
Arg1 plus court que Arg2	94 %	86 %	-0.6	0.20	0.00275
Arg2 indéfini	13 %	29 %	-0.22	0.10	0.0252
Arg2 nominal	27 %	48 %	-0.35	0.08	< 0.001
Arg2 pronominal	68 %	58 %			

On trouve plus fréquemment à l'oral des sujets courts (94 %) et pronominaux (68 %), et moins fréquemment des sujets nominaux (27 %) et indéfinis (13 %) par rapport à l'écrit. Ainsi, bien que l'alternance actif/passif suive les mêmes contraintes à l'oral et à l'écrit, les deux médiums diffèrent par le poids qu'ils associent à celles-ci. Prises ensemble, les différences du Tableau 7 suggèrent que les productions à l'oral respectent davantage la contrainte de codage harmonique (sujets courts et pronominaux, compléments indéfinis et nominaux). La contrainte de codage harmonique pouvant être interprétée comme une forme d'efficacité de codage des arguments (Haspelmath 2020 ; MacDonald 2013), nous avançons l'hypothèse que l'oral mette en jeu des productions syntaxiques plus efficaces que l'écrit. Cela peut s'expliquer par les conditions de traitement et de production de chaque médium : alors que l'écrit est un médium non interactif qui produit des énoncés statiques, l'oral est un processus dynamique limité dans le temps (Halliday 1989 ; Koch & Oesterreicher 2001). Avec moins de temps pour planifier les énoncés, les locuteurs à l'oral tendraient donc à choisir les codages les plus efficaces et rapides à traiter.

5. CONCLUSION

À travers cette étude, nous avons pu apporter une nouvelle description de l'alternance actif/passif à l'oral. L'utilisation des méthodes de modélisation statistique nous a permis de quantifier l'effet de facteurs variés, corroborant les études antérieures sur le passif. Nous avons ainsi rendu compte du caractère multifactoriel et probabiliste de l'alternance. De multiples contraintes de longueur croissante des constituants, de codage harmonique, de sémantique lexicale et d'amorçage sous-tendent les préférences syntaxiques des locuteurs et guident leur choix de construction. L'utilisation d'un actif ou d'un passif permet alors de maximiser le respect de ces contraintes. Soulignons que les résultats de notre étude montrent pour l'essentiel une unité de la construction passive à travers les médiums écrits et oraux. Enfin, nous avons abordé la question des différences entre écrit et oral dans le poids accordé à chaque contrainte, en avançant que l'oral semble être un médium qui privilégie davantage les codages harmoniques d'arguments, de façon à être plus efficace face aux limites temporelles inhérentes à sa production et son traitement. Cette dernière hypothèse devrait être testée

par l'étude d'autres alternances de construction ou d'autres genres textuels à l'oral et à l'écrit. À travers cette étude, nous avons aussi montré l'intérêt des méthodes quantitatives sur corpus pour décrire les phénomènes de préférence syntaxique ou de variation. Par l'exploration de nouveaux facteurs avec les mêmes méthodes, l'étude du passif pourrait également être conduite plus avant.

Références bibliographiques

[CARET] KUHN M. (2008), "Building predictive models in R using the caret package", *Journal of Statistical Software* 28 (5), 1-26.

[CFPP2000] *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000*, CLESTHIA & Université Paris 3 – Sorbonne Nouvelle. [http://cfpp2000.univ-paris3.fr/]
BRANCA-ROSOFF S. *et alii* (2012), « Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000) ». [en ligne].

[C-ORAL-ROM] *Corpus Oral de Langues Romanes*, Université de Florence.
CRESTI E. & MONEGLIA M. (eds.) (2005), *C-Oral-Rom: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam, John Benjamins.

[CRFP] *Corpus de Référence du Français Parlé*, dirigé par J. Véronis, DELIC (Université d'Aix-Marseille & Délégation à la Langue Française, France). [https://www.projet-orfeo.fr/corpus-source/]
DELIC (2004), « Autour du *Corpus de référence du français parlé* », *Recherches sur le français parlé* 18, 11-42.

[FTB] *French Treebank. Corpus de référence pour le français : ressource lexicale et syntaxique richement annotée pour les linguistes, utilisable en TAL*, LLF, IUF, CNRS, CNRTL. [http://ftb.linguist.univ-paris-diderot.fr/]
ABEILLÉ A., CLÉMENT L. & LIÉGEOIS L. (2019), « Un corpus annoté pour le français : le French Treebank », *TAL* 60, 19-43.

[LME4] BATES D. *et alii* ([2014] 2015), "Fitting linear mixed-effects models using lme4", *Journal of Statistical Software* 67 (1), 1-48. [arXiv:1406.5823v]

AISSEN J. (1999), "Markedness and subject choice in optimality theory", *Natural Language & Linguistic Theory* 17, 673-711.

AISSEN J. (2003), "Differential object marking : Iconicity vs. economy", *Natural Language and Linguistic Theory* 21, 435-483.

BÎLBÎIE G., FAGHIRI P. & THUILIER J. (2021), « Syntaxe quantitative et expérimentale : objets et méthodes », *Langages* 223, 7-24.

BLANCHE-BENVENISTE C. (2000), « Analyse de deux types de passifs dans les productions de français parlé », *Études romanes* 45, 303-319.

BOCK J. K. (1986), "Syntactic persistence in language production", *Cognitive Psychology* 18 (3), 355-387.

BRESNAN J. & FORD M. (2010), "Predicting syntax: Processing dative constructions in American and Australian varieties of English", *Language* 86 (1), 168-213.

BRESNAN J., DINGARE S. & MANNING C. D. (2001), "Soft constraints mirror hard constraints: Voice and person in English and Lummi", in M. Butt & T. Holloway King (eds.), *Proceedings of the LFG 01 Conference*, (Hong Kong), Stanford, CSLI Publications, 13-32.

CANDITO M. & SEDDAH D. (2012), « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », dans G. Antoniadis, H. Blanchon & G. Sérasset

(éds), *Actes de la Conférence conjointe JEP-TALN-RECITAL 2012* (Grenoble, France), vol. 2, AFCP & ATALA, 321-334.

DA CUNHA Y. & ABEILLÉ A. (2020), « L'alternance actif/passif en français : une étude statistique sur corpus écrit », *Discours* 27, 10956.

DEBAISIEUX J.-M. & BENZITOUN C. (éds) (2020), *Langages* nº 219 : *Orféo : un corpus et une plateforme pour l'étude du français contemporain*, Malakoff, Armand Colin/Dunod.

DESCLÉS J.-P. & GUENTCHÉVA Z. (1993), « Le passif dans le système des voix du français », *Langages* 109, 73-102.

DRUETTA R. (2020), « Le passif à l'oral : phénoménologie et propriétés aspectuelles dans OFROM », *Studia linguistica romanica* 4, 150-174.

DUBOIS J. & DUBOIS-CHARLIER F. (1997), *Les verbes français*, Paris, Larousse-Bordas.

EITELMANN M. (2016), "Support for end-weight as a determinant of linguistic variation and change", *English Language & Linguistics* 20 (3), 395-420.

ESTIVAL D. (1985), "Syntactic priming of the passive in English", *Text – Interdisciplinary Journal for the Study of Discourse* 5 (1-2), 7-22.

HALLIDAY M.A.K. (1989), *Spoken and Written Language*, Oxford, Oxford University Press.

HAMMA B. (2015), « Agent passif en PAR *versus* sujet actif : les dessous d'un ‹contraste› », *Revue de Sémantique et Pragmatique* 37, 61-83.

HAMMA B., TARDIF A. & BADIN F. (2017), « Le passif à l'oral », Fiche publiée dans *Français contemporain vernaculaire – FRACOV*. [en ligne]

HASPELMATH M. (2020), "Role-reference associations and the explanation of argument coding splits", *Linguistics* 59 (1), 123-174.

HENLEY N. M., MILLIER M. & BAEZLEY J. A. (1995), "Syntax, semantics and sexual violence: Agency and the passive voice", *Journal of Language and Social Psychology* 14 (1-2), 60-84.

HUNDT M., RÖTHLISBERGER M. & SEOANE E. (2018), "Predicting voice alternation across academic Englishes", *Corpus Linguistics and Linguistic Theory* 17 (1), 189-222.

KAHANE S. *et alii* (2017), « Annotation micro- et macrosyntaxique manuelle et automatique de français parlé », *Journée Floral* 4. [en ligne]

KOCH P. & OESTERREICHER W. (2001), « Langage oral et langage écrit », *in* G. Holtus, M. Metzeltin & C. Schmitt (eds), *Lexikon der Romanistischen Linguistik*, Band I, 2: *Methodologie*, Tübingen, Max Niemeyer Verlag, 584-627.

LABOV W. (1972), *Sociolinguistic Patterns*, Philadelphia, University of Pennsylvania Press.

LACHERET A. *et alii* (2014), « Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé », dans F. Neveu *et alii* (éds), *4e Congrès Mondial de Linguistique Française – CMLF'14* (Berlin, Allemagne), SHS Web of Conferences 8, Les Ulis, EDP Sciences, 2675-2689.

MACDONALD M. C. (2013), "How language production shapes language form and comprehension", *Frontiers in Psychology* 4, 226. [en ligne]

POIRET R. & LIU H. (2020), "Some quantitative aspects of written and spoken French based on syntactically annotated corpora", *Journal of French Language Studies* 30 (3), 355-380.

ROLAND D., DICK F. & ELMAN J. L. (2007), "Frequency of basic English grammatical structures: A corpus analysis", *Journal of Memory and Language* 57 (3), 348-379.

STALLINGS L. M., MACDONALD M. C. & O'SEAGHDHA P. G. (1998), "Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift", *Journal of Memory and Language* 39 (3), 392-417.

SZMRECSANYI B. (2016), "About text frequencies in historical linguistics: Disentangling environmental and grammatical change", *Corpus Linguistics and Linguistic Theory* 12 (1), 153-171.

SZMRECSANYI B. *et alii* (2017), "Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English", *Glossa. A Journal of General Linguistics* 2 (1), 86.

THUILIER J. (2012a), « Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux », dans F. Neveu *et alii* (éds), *3e Congrès Mondial de Linguistique Française – CMLF'12* (Lyon, France), SHS Web of Conferences 1, Les Ulis, EDP Sciences, 2451-2469.

THUILIER J. (2012b), *Contraintes préférentielles et ordre des mots en français*, Thèse de l'Université Paris-Diderot – Paris VII.

THUILIER J. *et alii* (2021), "Word order in French: The role of animacy", *Glossa. A Journal of General Linguistics* 6 (1), 55.

WASOW T. (2002), *Postverbal Behavior*, Stanford (CA), CSLI Publications.

WEINER E. J. & LABOV W. (1983), "Constraints on the agentless passive", *Journal of Linguistics* 19 (1), 29-58.

ZAENEN A. *et alii* (2004), "Animacy encoding in English: Why and how", *Proceedings of the Workshop on Discourse Annotation* (Barcelona, Spain), Stroudsburg (PA), Association for Computational Linguistics, 118-125.

ZRIBI-HERTZ A. (1982), « La construction ‹se-moyen› du français et son statut dans le triangle moyen-passif-réfléchi », *Lingvisticae Investigationes* 6 (2), 345-401.

ZRIBI-HERTZ A. (2008), « Le médiopassif à SN préverbal en français : pour une approche multifactorielle », dans J. Durand, B. Habert & B. Laks (éds), *Congrès Mondial de Linguistique Française – CMLF'08* (Paris, France), Les Ulis, EDP Sciences & Institut de Linguistique Française, 2645-2662.