Les formes de sujet

Cette étude consiste à examiner les formes des sujets de chaque construction verbale, puisqu'une variété de natures de mot peuvent occuper cette fonction: les substantifs et, à l'oral, surtout les pronoms (personnels, démonstratifs, relatifs). Pour procéder de manière fiable à ces comptages, on convient de ne retenir pour cette étude que les "vraies" constructions verbales, c'est-à-dire qu'on écarte systématiquement les constructions incomplètes (dans les cas de bafouillage) et les formes qui n'ont pas une valeur de verbe, comme *c'est-à-dire* ou *tu vois*, *tu sais* (pour les locuteurs qui les emploient de façon récurrente, comme un tic de langage). De même, dans des construction comme *ce qui m'embête c'est qu'il n'a rien dit*, on ne compte qu'un seul sujet sous dispositif, et non deux sujets pronoms. Toutes ces mesures ont pour but de ne pas sur-représenter la classe des pronoms clitiques, qui est déjà la plus nombreuse dans les corpus de français parlé.

Dans le tableau suivant, on voit pour chaque corpus, la distribution entre sujets lexicaux, pronoms clitiques (séparés en 2 catégories: 3e personne [*il, elle, ce, ça, on*] et 1e/2e personnes [*je, tu, nous, vous*]) et autres, c'est-à-dire pronoms relatifs, "gros pronoms" (*personne, rien, certains…*) et sujets sous dispositif.

	TOTAL SUJETS	sujet lexical	clitiques 3e personne	clitiques 1e/2e pers.	autres
Couches	85	10	55	6	14
Lentilles	183	30	68	70	15
Accident	262	13	86	131	32
Adoucisseur	342	45	187	82	28
Éducation	75	5	32	30	8
Guerre	126	1	73	41	11
Résidence	96	2	79	6	9
Oeufs	120	24	78	4	14
Vie paris	198	2	124	58	14
Chasse	94	1	64	26	3
Agence	311	17	133	149	12
Belfast	259	2	140	109	8
TOTAL	2151	152	1119	712	168

On voit qu'en moyenne (sur la totalité des 2150 constructions verbales du corpus entier), les sujets lexicaux sont les plus rares: ils représentent seulement 7% des sujets, alors que les pronoms clitiques se rencontrent dans 85% des cas.

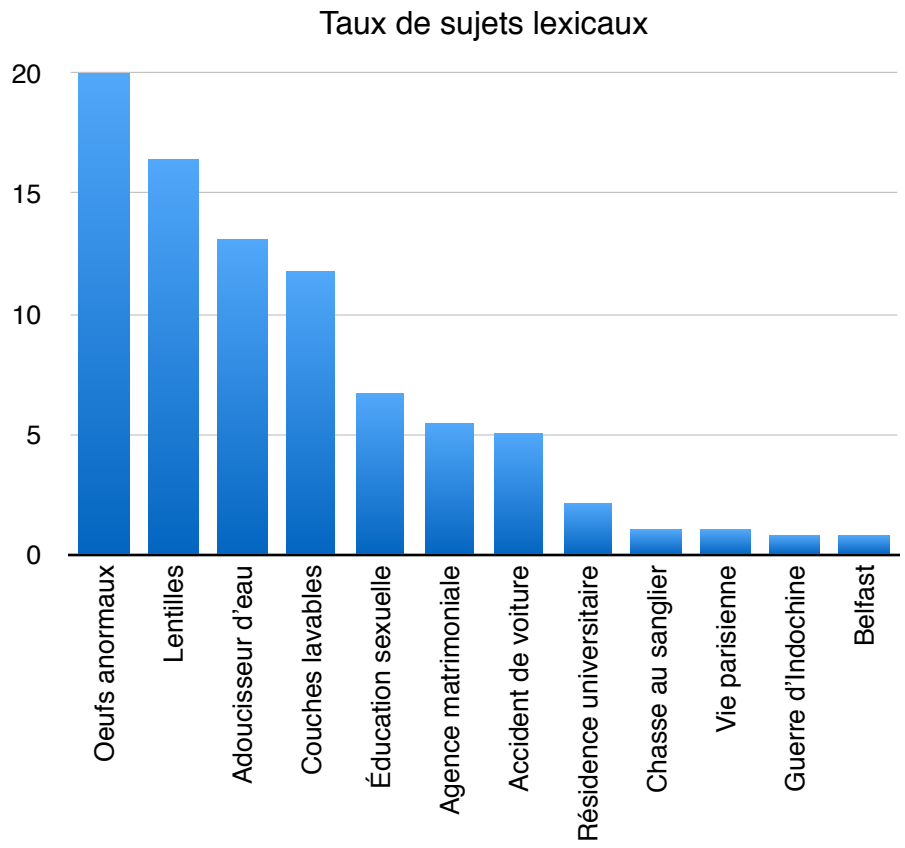
Cette relative rareté des sujets lexicaux en fait un bon indicateur du registre de langue des différents corpus. Les taux élevés de sujets lexicaux correspondent à un langage plus soigné, plus proche de l'écrit:

> donc dans un troisième temps **l'eau** va s'infiltrer à travers les couches du sol (Adoucisseur d'eau)

En français de conversation, on aurait tendance à éviter les sujets lexicaux soit par un double-marquage, soit par un dispositif:

> **l'eau elle** va s'infiltrer
> **il y a de l'eau qui** va s'infiltrer

Les formes significatives pour évaluer le registre de langue sont donc: les sujets lexicaux (figure 1), les sujets clitiques (figures 2, 3) et les double-marquages (figure 4).

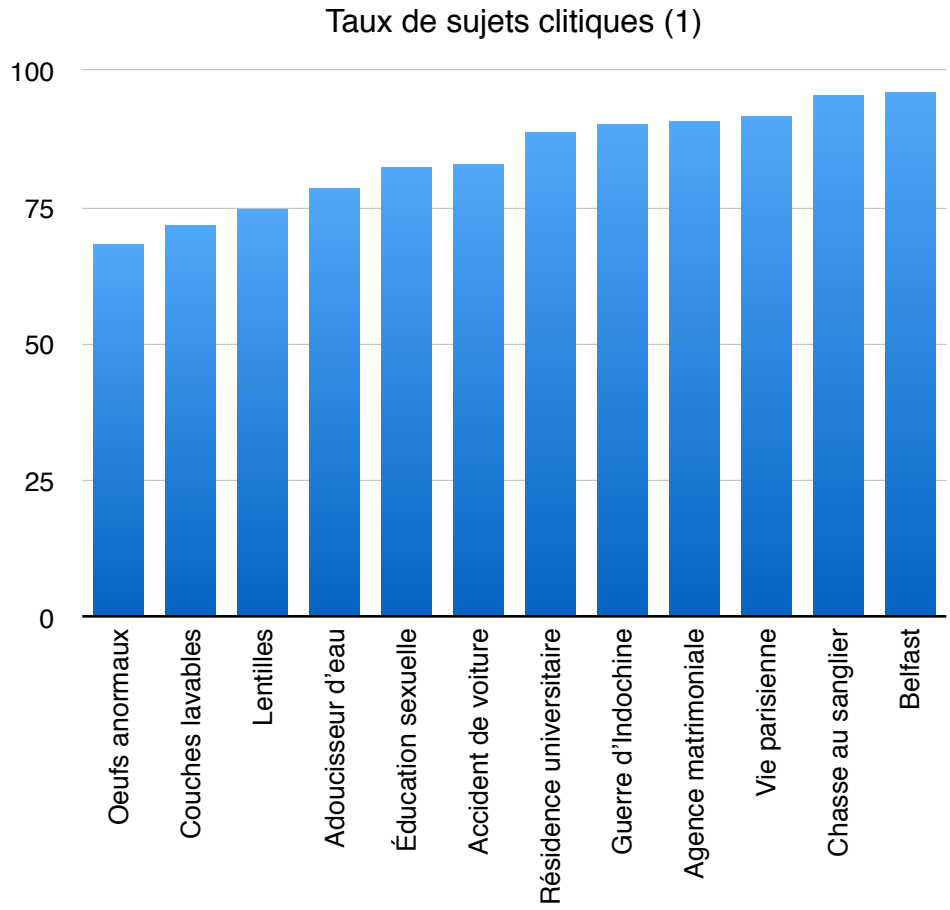


On voit clairement 3 groupes de corpus: à gauche, 4 corpus avec un taux élevé de sujets lexicaux (entre 12 et 20%); au milieu, 3 corpus moyens (entre 5 et 7%) et à droite, 5 corpus avec très peu de sujets lexicaux (entre 0 et 2%). Excepté le corpus *Oeufs anormaux*, qui a le taux de sujets lexicaux le plus élevé tout en étant d'un style peu formel suivant les autres marqueurs (aucun *ne* de négation et peu de liaisons), les 3 groupes sont consistants avec les registres de langue déjà identifiés: la langue formelle et soignée pour *Lentilles*, *Adoucisseur* et *Couches lavables*; la

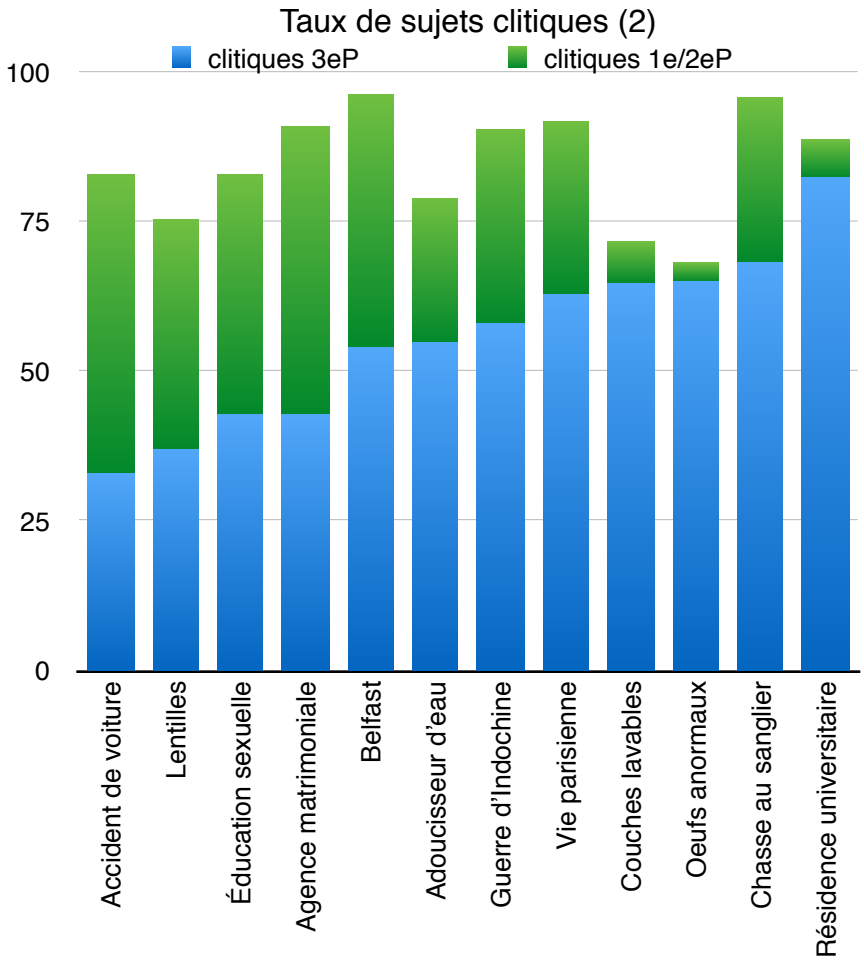
langue familière pour *Belfast*, *Guerre d'Indochine*, *Vie parisienne, Chasse au sanglier* et *Résidence universitaire*.

Le corpus *Oeufs anormaux* est un peu paradoxal. Son haut taux de sujets lexicaux ne proviendrait pas d'un registre particulièrement soigné, mais plutôt du fait qu'il s'agit d'un corpus d'explication, où le locuteur donne de l'information objective sur les poules:

C'est un phénomène qui va se produire euh souvent lorsque **la poule** est très jeune, quand que **la poulette** commence à pondre des œufs, vers l'âge - - de seize semaines, **son système** est comme pas synchronisé encore, pis ses premiers x œufs que **la poule** va pondre, il y aura pas de jaune à l'intérieur.

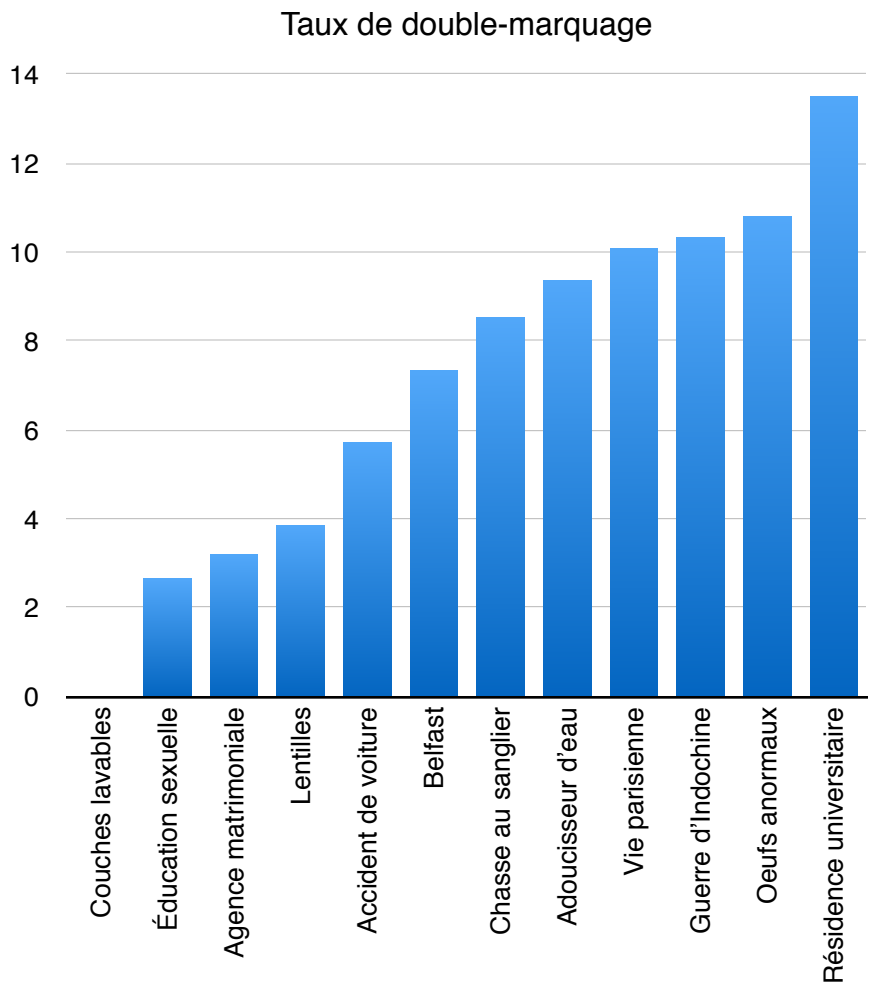


Ce tableau, qui présente les différents taux de clitiques tous clitiques confondus, ne nous apprend pas grand-chose. À droite, les taux les plus élevés correspondent aux corpus les plus "conversationnels", alors qu'à gauche, on a des corpus plus soignés ou moins "conversationnels". Pour une étude plus précise, il est intéressant de distinguer entre clitiques de 3e personne et clitiques du discours (*je, tu, nous, vous*):



On voit que le corpus *Oeufs anormaux* a le taux le plus bas de clitiques de 1e et 2e personnes, ce qui confirme son style ”non-conversationnel”. À l’opposé, les corpus qui en ont le plus sont les récits auto-biographiques (les cinq corpus à gauche), forcément centrés sur le *je*, surtout dans le cas d’aventures ou d’anecdotes (*Accident de voiture, Agence matrimoniale, Éducation sexuelle*). Notons tout de même que ce n’est pas le cas pour tous les récits de vie: *Guerre d’Indochine* et *Vie parisienne* ont un taux modéré de clitiques de 1e et 2e personnes, comparable à celui de corpus d’explication comme *Adoucisseur d’eau* et *Chasse au sanglier.* On pourrait parler alors de récits de vie plus factuels, par opposition à ceux qui sont centrés sur le *je*. On voit donc que finalement, l’analyse des sujets est un outil qui ne révèle pas seulement les registres de langue, mais aussi les types de discours.

Le dernier type de sujet à observer est le double-marquage, un cas particulier de clitiques de 3e personne couplés avec une forme lexicale, de type ”l’eau elle va s’infiltrer”. Ces formes sont moins nombreuses dans les corpus, mais on espère qu’elles seront un meilleur indicateur du niveau de langue.



Le taux de double-marquage du sujet varie de 0% à 13.5%, ce qui est considérable. Les corpus à taux nul ou très bas (au-dessous de 4%) sont des corpus formels, excepté *Agence matrimoniale*, un corpus de français assez familier où la locutrice ne fait pas d'effort particulier pour les liaisons ou les ne de négation. Les taux élevés (au-dessus de 8%), correspondent en gros aux corpus plus familiers, excepté *Adoucisseur d'eau* qui a par ailleurs quelques caractéristiques de français formel. Notons aussi le corpus *Oeufs anormaux*, qui présente le paradoxe d'avoir le plus haut taux de sujets lexicaux et le 2e plus taux de double-marquage!

Aucune de ces études statistiques n'est jamais un outil parfait. Il y a toujours des exceptions, des bizarreries, des corpus hors norme… Les statistiques peuvent tout au plus nous aider à trouver quelques tendances, mais elles ne parviennent pas à caractériser de façon définitive le registre de langue d'un corpus.