Effects of literary form on grammatical changes: A treebank study

This paper investigates the effects of the text form (prose vs. verse) on diachronic grammatical changes in Medieval French using parsed treebanks [5] and [6] (∼ 1 mln words annotated with Penn scheme).

Despite the common intuition that the prose is somehow more “advanced” than the verse contemporary to it with respect to grammatical changes, in the absence of quantificational evaluations, the magnitude of the difference has remained unknown. At the same time, the prevalence of verse in the earliest periods of documented French (i.e. X–XII c.) results in a strong and unavoidable correlation between time and form, which potentially undermines the results of the studies attempting to formally model Medieval French evolution. This study contributes to solving this problem by comparing two historical changes, the loss of pro-drop and that of OV \textit{finite} order, across text forms.

We model the loss of pro-drop using logistic regression on a sample of finite clauses with either an overt pronominal or null subject (total of 95,574), excluding relatives, imperatives, and \textit{wh}-questions because of their idiosyncratic subject syntax. Our model 
\[
P(Subject = \text{overt} | \text{Date}) = \frac{e^{\alpha + \beta \times \text{Date}}}{1 + e^{\alpha + \beta \times \text{Date}}}
\]
has $\alpha=-1.889$ and $\beta=0.0027$ for prose and $\alpha=-2.083$ and $\beta=0.0021$ for verse (Fig. 1). That is, the rates of change ($\beta$) are nearly identical, whereas the difference in intercepts means that the verse is more “conservative”, which is in line with the traditional intuition. This also points towards a possible extension of Kroch’s (1989) Constant Rate Hypothesis according to which a given grammatical change has the same rate across (grammatical) contexts unto different text forms. This, however, does not hold for the loss of OV (Fig. 2, 
\[
P(\text{Position} = \text{OV} | \text{Date}) = \frac{e^{\alpha + \beta \times \text{Date}}}{1 + e^{\alpha + \beta \times \text{Date}}}
\]
with $\alpha=7.682$ and $\beta=-0.008$ for the prose and $\alpha=0.276$ and $\beta=-0.0007$ for the verse), which appears to be almost constant for the verse. This suggests that the verse is to be treated separately at least for the evolution of OV, which has not been the practice yet (cf. [7], [3], [2], [8], [4], [1]). The analysis of the effect (e.g. verse being an “artificial” syntactic register etc.) remains for future research.


