Fitting the Development of Periphrastic *do* in All Sentence Types

Relja Vulanović[1] and Harald Baayen[2]
[1]Department of Mathematical Sciences, Kent State University, Stark Campus, 6000 Frank Ave. NW, Canton, Ohio, USA, e-mail: rvulanovic@stark.kent.edu and
[2]Max Planck Institute of Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, The Netherlands, e-mail: baayen@mpi.nl

ABSTRACT

The historical development of periphrastic-*do* constructions in English is considered in five different sentence types. This syntactic change is viewed in all contexts as a two-stage change, which motivates the choice of fitting curves to the data collected by Ellegård. Very good fits are obtained simultaneously for all sentence types.

1. INTRODUCTION

The historical development of periphrastic *do* in different types of English sentences is well-documented in (Ellegård, 1953), where the periphrastic-*do* construction is analyzed in 107 texts between 1390 and 1710. Ellegård's examples illustrating this syntactic change can be found also in (Kroch, 1989a, 1989b), (Ogura, 1993), and (Vulanović, 2005, to appear). These papers use Ellegård's data to discuss the change further. Kroch (1989a, 1989b) and Ogura (1993) give plausible linguistic explanations of the development of periphrastic *do* in different types of sentences. Vulanović (2005) uses his grammar efficiency model to confirm Ellegård's hypothesis that emphatic *do* influenced the development in affirmative declarative sentences. In this type of sentences, periphrastic *do* initially increases up to about 10% and then decreases and almost disappears. This behavior is different from what can be observed in other sentence types (negative declaratives, negative imperatives, and affirmative and negative questions), where the data show a gradual S-shaped increase in the proportion of sentences with periphrastic *do*. Because of the S shape, the simple logistic curve can be used to fit the data and this is done in (Kroch, 1989a, 1989b) (see (Kroch, 2001) as well) and (Ogura, 1993). Kroch and Ogura do not consider any other fitting curves and do not provide any fit for the affirmative-declarative data. This is done in (Vulanović, to appear), where two different approaches are successfully applied. Both approaches are based on some appropriate modifications of the simple logistic curve. This curve solves the logistic differential equation with a constant coefficient k. When this coefficient is replaced with a function of time, $k(t)$, the resulting solution is a generalized logistic curve. $k(t)$ is a linear function in (Altmann, 1983) and (Best et al., 1990), which is suitable for fitting reversible linguistic changes, and is therefore used in (Vulanović, to appear) as well. The other approach in (Vulanović, to appear) starts from the logistic differential equation with a piecewise constant function $k(t)$. This gives a curve which is a combination of two simple logistics, an increasing S-shaped curve followed by a decreasing one. When the curves are linearized, the method is equivalent to the linear regression with an unknown changeover point (Seber, 1977:p. 208). The same kind of combination of two logistic curves is used also in (Imsiepen, 1983) to model the development of *e*-epithesis in strong German verbs.

Since the periphrastic-*do* data in different contexts have been fitted separately so far, our interest here is to find a unifying fit. One of the methods we use is the combination of two simple logistics, applied this time to all sentence types, not just to affirmative declaratives like in (Vulanović, to appear). We investigate also another possible class of fitting curves, those arising from the logistic differential equation with a quadratic coefficient $k(t)$. This too is combined with a changeover point. We show that these approaches provide effective fits to all periphrastic-*do* data simultaneously.

In section 2, we present Ellegård's data and discuss the generalized logistic differential equation and its solution. Section 3 contains the results of curve-fitting. We finish with a brief conclusion.

2. ELLEGÅRD'S DATA AND THE GENERALIZED LOGISTIC CURVE

Table 1 is based on the data from Table 7 in (Ellegård, 1953). Like in (Vulanović, to appear), the Ellegård's thirteen periods are reduced to eleven by merging together the original first and second periods, as well as the last two ones. This is done because Ellegård considers fewer texts in these periods. Like Ogura (1993), we include negative imperative sentences in the discussion. They are not considered in (Kroch, 1989a, 1989b). Kroch and Ogura distinguish between different types of affirmative questions, but there is no need to follow their suit here.

Table 1 is represented graphically in Figure 1. The time coordinates of the plotted points are the midpoints of each of the eleven periods. All the graphs and statistical calculations in this paper are done in R, a public-domain statistical programming environment.

Period	AD	ND	AQ	NQ	NI
1390–1425	0.0003	0.	0.	0.118	0.
1425–1475	0.0027	0.012	0.042	0.080	0.011
1475–1500	0.0178	0.048	0.070	0.111	0.
1500–1525	0.0138	0.078	0.227	0.590	0.120
1525–1535	0.0263	0.137	0.324	0.607	0.
1535–1550	0.0815	0.279	0.449	0.750	0.
1550–1575	0.0932	0.380	0.563	0.854	0.093
1575–1600	0.0634	0.238	0.603	0.648	0.064
1600–1625	0.0304	0.367	0.692	0.937	0.353
1625–1650	0.0294	0.317	0.829	0.842	0.238
1650–1710	0.0136	0.544	0.825	0.941	0.738

Table 1: Proportion of periphrastic *do* in all sentence types (AD = affirmative declarative, ND = negative declarative, AQ = affirmative question, NQ = negative question, NI = negative imperative)

It can be observed that the proportion of periphrastic-*do* constructions shows a generally increasing trend in all non-AD sentences, but there is a slow-down, or in most cases even a decrease, around 1560, exactly at the same time when periphrastic *do* started its definite decline in affirmative declaratives. The change in each sentence type can therefore be viewed as a two-stage change. This motivates some special choices of the coefficient $k(t)$ in the generalized logistic differential equation. This equation is

$$\frac{dp(t)}{dt} = k(t)p(t)[m - p(t)], \tag{1}$$

where p is the proportion of the nascent linguistic form (or a construction), which is a function of time t, $t \geq 0$, and m is a constant, $0 < m \leq 1$, indicating the maximum value that p approaches. Since $m - p$ is the proportion of the old linguistic form which is being replaced with the new one, equation (1) means that the rate of change of p is directly proportional to the amount of both new and old forms and their interaction. The solution of (1) is

$$p(t) = \frac{m}{1 + \exp\left[-K(t)\right]}, \quad K(t) = m \int k(t)\, dt. \tag{2}$$

The linguistic change described above is known as the Piotrowski Law or

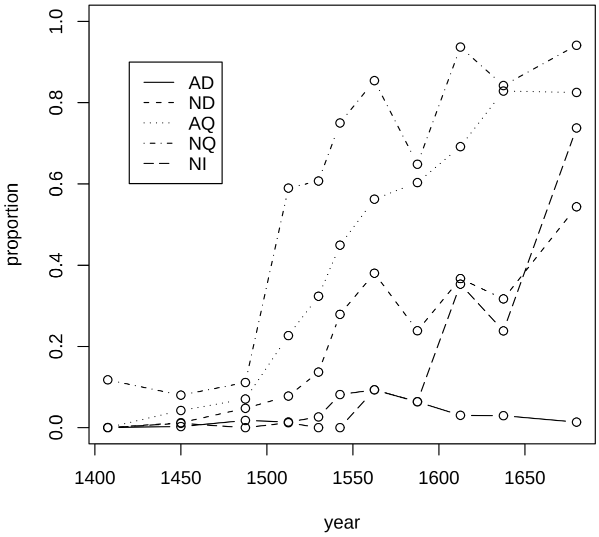

Figure 1: Table 1 data represented graphically

the Piotrowski-Altmann Law. Altmann (1983) distinguishes between three kinds of change:

(i) complete change when $m = 1$ and $k(t) =$ const.,

(ii) partial change when $m < 1$ and $k(t) =$ const., and

(iii) reversible change when $k(t)$ is a linear function of t.

Cases (i) and (ii) lead to the simple logistic curve and case (iii) to a generalized one. Case (i) suffices for fitting the non-AD data (Kroch, 1989a, 1989b; Ogura, 1993), whereas (iii) with $m \approx 0.1$ is suitable for the change in AD sentences (Vulanović, to appear). Since our interest here is to use the same class of functions to fit all sentence types, we keep the two-stage nature of all periphrastic-*do* changes in mind and consider this as a fourth kind of linguistic change. A two-stage change can be reversible but it also covers the case of two increases separated by a short period of stagnation or decline. Therefore, the piecewise-constant choice for $k(t)$, which has already been considered in (Vulanović, to appear) for AD sentences, is a natural choice for all two-stage changes. In this approach,

$$k(t) = \begin{cases} k_\ell & \text{if } t \le T \\ k_r & \text{if } t > T \end{cases}, \tag{3}$$

where k_ℓ and k_r are two constants and T is the time coordinate of the point where one simple logistic curve is replaced with another.

Another choice of $k(t)$ for a two-stage change is that of a quadratic function,

$$k(t) = 3At^2 + 2Bt + C, \tag{4}$$

so that (2) becomes

$$p(t) = \frac{1}{1 + \exp[-(At^3 + Bt^2 + Ct + D)]}, \tag{5}$$

where D is an integration constant. The choice in (4) is attractive because of the following possible interpretation. If we assume that $A > 0$ and that the parabola (4) has two t-intercepts, then $k(t)$ changes its sign from positive to negative and then again to positive. Because of this, it follows form (2) that the sign of $dp(t)/dt$ changes in the same way (assuming the solution $p(t)$ stays between 0 and m). We can therefore expect that $p(t)$ changes from increasing to decreasing and then back to increasing. This is exactly how most of the periphrastic-*do* data look like.

3. THE FITS

We consider in this section two main fitting methods based on (3) and (4) respectively, with an additional variation of the latter. Within each method, all fitting curves are obtained simultaneously with $m = 1$ using the generalized-linear-model function in R with the binomial family linked to the logit function. The dependent variable is a two-column matrix whose columns contain proportions of sentences with and without *do* for each year and each sentence type.

Fit I. We use (3) with the changeover point T set at the seventh measurement point in time. This is somewhat different from what is done in (Vulanović, to appear), but is simpler. The changeover point is introduced via Indicator, a variable set equal to 0 for the seven initial values of t, and to 1 after that. We get a very good fit to the data with the coefficient of determination $R^2 = 0.96$ for the entire model. This value is included in

Fit	All	AD	ND	AQ	NQ	NI
I	0.96	**0.89**	0.91	0.96	0.93	0.91
II	0.95	**0.73**	0.91	0.99	0.88	0.87
IIa	0.97	0.92	0.94	0.98	0.94	**0.90**

Table 2: R^2 values (All = complete model; values for separate sentence types are obtained from the appropriate subsets of the data after the fit was found for the whole model)

	Df	Dev	Resid. Df	Resid. Dev	F
NULL			54	20932.3	
year	1	6557.3	53	14375.0	6557.27
type	4	8274.0	49	6101.0	2068.50
Indicator	1	1161.5	48	4939.6	1161.46
year:type	4	659.2	44	4280.4	164.79
type:Indicator	4	103.4	40	4177.0	25.86
year:Indicator	1	3236.1	39	940.9	3236.11

Table 3: F-tests for the predictors in Fit I; $\Pr(> F)$ <2.2e–16 in all cases (type = AD, ND, AQ, NQ, NI; Indicator = 0 for year $\leq$ 1562.5 and 1 otherwise)

Table 2 together with R^2-values for other fits and for all individual sentence types. The smallest R^2-value in each fit is boldfaced.

Table 3 presents the results of F-tests for the predictors in the model. All predictors are highly significant. A visual impression of the fit is shown in Figure 2. Each curve is a combination of two logistics of type (5) with $A = B = 0$. The coefficients C and D are given in Table 4. We emphasize again that the fit was obtained for all curves simultaneously.

Fit II. This fit is based on (4). Table 2 shows that the results are now worse, particularly for AD sentences. Because of this, we consider the following modification.

Fit IIa. A changeover point is introduced at the third measurement point

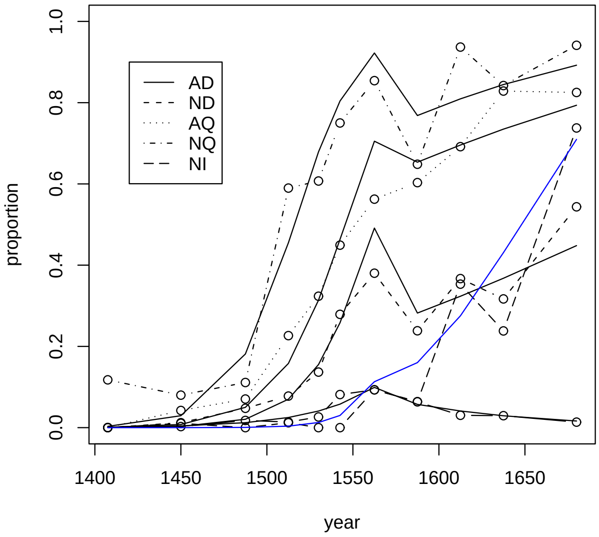

Figure 2: Fit I

in time because the data change very little up to that point. Like in Fit I, this is done via Indicator, which is this time set equal to 0 for the three initial values of t, and to 1 after that. Table 2 shows the best R^2-values. All predictors are again significant, as indicated in Table 5. The graph is presented in Figure 3 and the coefficients for each curve, consisting of two type (5) curves linked together, are given in Table 6. Like in Fit I, Indicator is used so that its two values affect only the coefficients C and D. No improvement is achieved when we let all coefficients change.

4. CONCLUSION

One of the contributions of this paper is the introduction of a new kind of linguistic change to complement Altmann's (1983) classification based on the logistic curve. To Altmann's complete, partial, and reversible changes, we add the two-stage change, inspired by how the graphs of periphrastic-*do* data look like in Figure 1. Around 1560, all sentence types except affirmative questions show a decline in the use of periphrastic *do*, from which only affirmative declaratives do not recover. The five developments cannot be

	Indicator = 0		Indicator = 1	
type	C	D	C	D
AD	2.92e–2	–47.78	–1.40e–2	19.47
ND	5.10e–2	–79.79	7.85e–3	–13.39
AQ	5.09e–2	–78.66	7.70e–3	–11.59
NQ	5.31e–2	–80.47	9.89e–3	–14.49
NI	7.08e–2	–112.60	2.76e–2	–45.44

Table 4: Fit I – coefficients C and D in formula (5) ($A = B = 0$)

described by one linguistic-change category of Altmann's, but they all fall within the new two-stage type.

The change in periphrastic *do* is a single syntactic change with five manifestations in different sentence types. It is therefore natural to fit all five developments simultaneously. We accomplish this by using curves (2) with $m = 1$ and with either piecewise linear or cubic functions $K(t)$, which are suitable for fitting two-stage linguistic changes. The resulting fits, obtained for all curves at the same time, are very good. By considering an overall model like this, we do not focus on individual curves and we avoid the danger of over-fitting them.

References

Altmann, G. (1983). Das Piotrowski–Gesetz und seine Verallgemeinerungen. In K.-H. Best & J. Kohlhase (Eds.), *Exacte Sprachwandelforschung* (pp. 54–90). Göttingen: Herodot.

Best, K.-H., Beöthy, E., & Altmann, G. (1990). Ein methodischer Beitrag zum Piotrowski-Gesetz. *Glottometrika*, *12*, 115–124.

Ellegård, A. (1953). *The auxiliary do: The establishment and regulation of its use in English*. Stockholm: Almquist & Wiksell.

Imsiepen, U. (1983). Die e-Epithese bei starken Verben im Deutschen. In K.-H. Best & J. Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 119-141). Göttingen: Herodot.

	Df	Dev	Resid. Df	Resid. Dev	F	Pr($> F$)
NULL			54	20932.3		
year	1	6557.3	53	14375.0	6557.27	$<$ 2.2e–16
year^2	1	3999.5	52	10375.6	3999.48	$<$ 2.2e–16
year^3	1	32.5	51	10343.0	32.51	1.2e–08
type	4	7864.7	47	2478.3	1966.18	$<$ 2.2e–16
Indicator	1	175.7	46	2302.6	175.72	$<$ 2.2e–16
year:type	4	931.6	42	1371.0	232.90	$<$ 2.2e–16
year^2:type	4	221.5	38	1149.5	55.38	$<$ 2.2e–16
year^3:type	4	58.8	34	1090.6	14.71	5.1e–12
type:Indicator	4	48.3	30	1042.3	12.08	8.1e–10
year:Indicator	1	486.6	29	555.7	486.65	$<$ 2.2e–16

Table 5: F-tests for the predictors in Fit IIa (type = AD, ND, AQ, NQ, NI; Indicator = 0 for year $\leq$ 1487.5 and 1 otherwise)

Kroch, A. S. (1989a). Function and grammar in the history of English: Periphrastic *do*. In R. W. Fasold & D. Schiffrin (Eds.), *Language change and variation* (pp. 133–172). Amsterdam: John Benjamins.

Kroch, A. S. (1989b). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*, 199–244.

Kroch, A.S. (2001). Syntactic change. In M. Baltin & C. Collins (Eds.), *The handbook of contemporary syntactic theory* (pp. 699–729). Oxford: Blackwell.

Ogura, M. (1993). The development of periphrastic *do* in English: A case of lexical diffusion in syntax. *Diachronica*, *10*, 51–85.

Seber, G.A.F. (1977). *Linear regression analysis*. New York: Wiley.

Vulanović, R. (2005). The rise and fall of periphrastic *do* in affirmative declaratives: A grammar efficiency model. *J. Quantitative Linguistics*, *12*, 1–28.

Vulanović, R. (to appear). Fitting periphrastic *do* in affirmative declaratives. (paper presented at QUALICO 2003), *J. Quantitative Linguistics*.

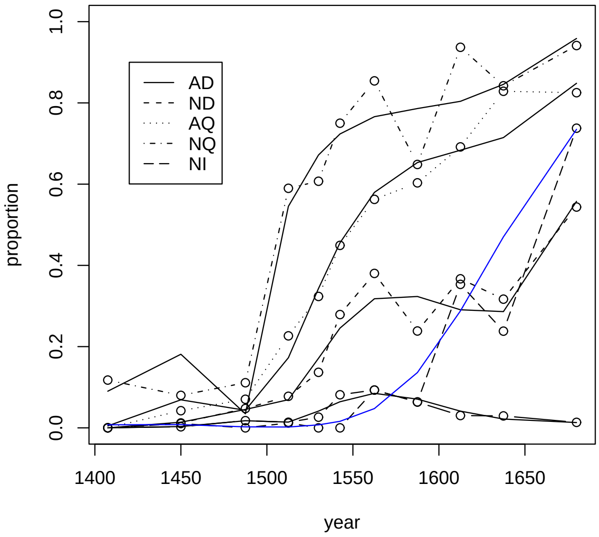

Figure 3: Fit IIa

type	A	B	Indicator = 0		Indicator = 1	
			C	D	C	D
AD	3.05e–6	–1.48e–2	23.83	–1.27e+4	23.97	–1.29e+4
ND	3.32e–6	–1.59e–2	25.39	–1.34e+4	25.53	–1.36e+4
AQ	1.98e–6	–9.57e–3	15.28	–8.08e+3	15.42	–8.29e+3
NQ	2.00e–6	–9.55e–3	15.04	–7.83e+3	15.17	–8.04e+3
NI	5.93e–7	–2.99e–3	4.92	–2.66e+3	5.06	–2.86e+3

Table 6: Fit IIa – coefficients in formula (5)