TWO CANDIDATE APPROACHES TO TEXT SEQUENCING:[1] AN ADDENDUM TO "THE DIACHRONY DEBATE: A TUTORIAL ON METHODS"

A. Dean Forbes

Department of Hebrew
University of the Free State
E-mail: adforbes@post.harvard.edu

(Received 02/05/2017; accepted 04/07/2017)

DOI: https://doi.org/10.25159/1013-8471/3679

ABSTRACT

In a recent essay published in this journal, I illustrated the limitations one may encounter when sequencing texts temporally using s-curve analysis. I also introduced seriation, a more reliable method for temporal ordering much used in both archaeology and computational biology. Lacking independently ordered Biblical Hebrew (BH) data to assess the potential power of seriation in the context of diachronic studies, I used classic Middle English data originally compiled by Ellegård. In this addendum, I reintroduce and extend s-curve analysis, applying it to one rather noisy feature of Middle English. My results support Holmstedt's assertion that s-curve analysis can be a useful diagnostic tool in diachronic studies. Upon quantitative comparison, however, the *five-feature* seriation results derived in my former paper are found to be seven times more accurate than the *single-feature* s-curve results presented here.

S-CURVES AND SINGLE-FEATURE ANALYSIS

Using proxy data from Middle English (ME),[2] I show how: 1) sequenced single-feature data are approximated by a traditional s-curve; and 2) unsequenced single-feature data may be fitted to an assumed s-curve.

[1] Presented at a 2016 International Syriac Language Project session in Stellenbosch, South Africa. My thanks to Professors Zevit, Naudé, Miller-Naudé, and Holmstedt for their helpful comments. All remaining errors and obscurities are my sole responsibility.

[2] Made advisable by our present lack of widely accepted independently sequenced Biblical Hebrew datasets.

The s-curves of historical linguistics are idealisations

Linguists with a quantitative bent often approach historical data by way of innovation theory as exemplified by s-curve analysis.[3] The basic idea is that, over time, the relative frequency of occurrence of a linguistic innovation tends to obey an s-curve of the sort shown in Figure 1.

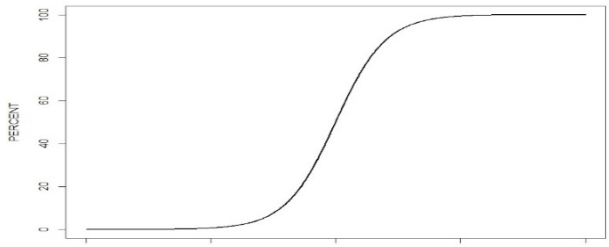

Figure 1. Idealised S-curve

In our context, the horizontal axis is *time*. Early on, innovation adopters are few and slowly increasing in number. During the middle phase, adopter numbers snowball as the innovation escalates. In the final phase, stragglers adopt the innovation, and the relative frequency of usage approaches an upper limit.

In three respects, the s-curve in Figure 1 is an idealisation in that actual innovation data: 1) exhibit fluctuations; 2) need not achieve and maintain 100% saturation;[4] and 3) may achieve dominance, only subsequently to peter out.[5]

Fitting an s-curve to sequenced feature data

Historical linguists typically wish to fit an s-curve to a dated linguistic feature, thereby quantitating the extent to which standard innovation theory approximates the observed data. This sub-section shows one linguistic feature from Ellegård's data (Ellegård 1953:161, Table 7, columns 8 and 9) and the associated estimated s-curve. The feature

[3] For a BH perspective, see Forbes (2017:§2.5).

[4] Consider, for example, *alternants* in language. In English, both *Mary gave the book to John* and *Mary gave John the book* occur. Neither alternant has driven the other out of English (yet?).

[5] Cassette tape, anyone? For an instance from BH, see Forbes (2017:§2.5.2).

tracks the relative incidence in ME of do-forms (do, does, did) in negative questions over three centuries.

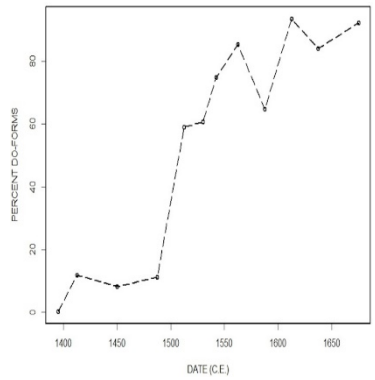

Figure 2. Incidence of Do-forms in Negative Questions

Figure 2 shows the fluctuating raw data. Ellegård ascribed the fluctuations, in general, to small sample sizes but speculated that the dip in the second half of the sixteenth century was due to transient changes in ME (Ellegård 1953:162–163).

Figure 3 shows a fitted s-curve optimally overlaying the observed data. Note: 1) the s-curve is a smoothed fit to the jagged observed data; and 2) the estimated upper limit is not 100% but around 84%, there being little data present at the dates tallied to enforce a 100% upper limit.

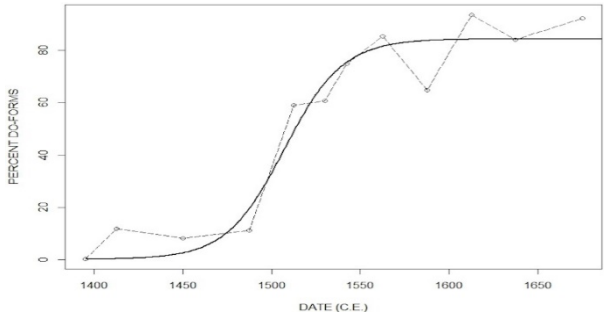

Figure 3. S-curve Fitted to Observations

Fitting an undated feature to an s-curve

Use of idealised s-curves in diachrony studies

In BH diachrony studies, we have various corpora exhibiting various linguistic features. We do not have associated dates or date ranges. Analysis seeks to infer these.

Holmstedt (2012:113–119) studied the counts of -שׁ as a percentage of (-שׁ + אשׁר) counts in BH as follows:

1. He assumed an idealised s-curve shape exhibiting eventual full and permanent 100% saturation.
2. He ordered the incidence data from smallest to largest, necessarily ignoring fluctuation effects.
3. He plotted the ordered data on an idealised s-curve.

According to Holmstedt (2016:242, n. 41), the "use of the idealized S-curve does not represent a statistical analysis per se but [is] an approximating diagnostic tool [useful] to check the plausibility of a diachronic analysis." Is this a reasonable stance?[6]

Precursor Figure 4

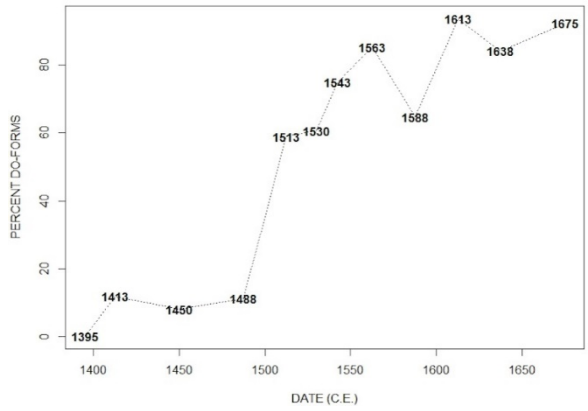

Figure 4. Mid-range Dates for Neg. Quest. Do-forms

[6] According to Rezetko and Young (2014:234, n. 82): "It would be a fundamental misuse of the S-curve to try to use it to sequence or date linguistic phenomena or the writings containing them when the dates of origin of those writings have not been determined independently before hand."

Figure 4 is a variant of Figure 2 wherein the little circles in Figure 2 have been replaced by centred “mid-epoch” date labels. The (fluctuating) heights of the year labels reveal their associated do-form percentages. Moving across the plot, the dates increase ceaselessly, but the associated percentages fluctuate.

Sequencing the data to yield Figure 5

If we sort the data in Figure 4 so that the percentages, rather than the dates, increase ceaselessly, Figure 5 results.

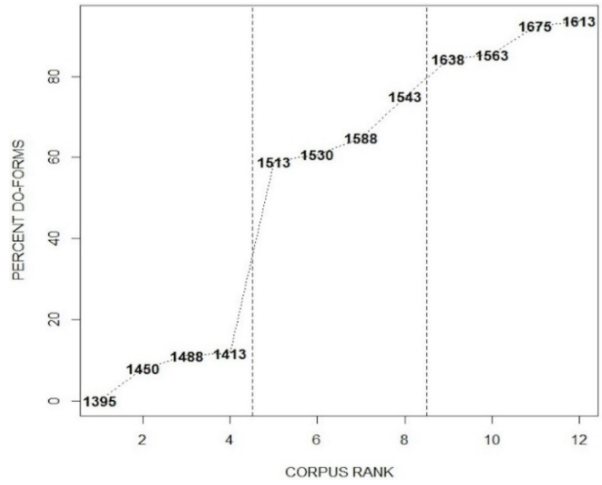

Figure 5. Increasing Percentages for Neg. Quest. Do-form

This is analogous to what Holmstedt did with his data, except that here we know the true dates. The dates do not increase ceaselessly, but do sequence in promising ways. The leftmost quartet, although out of order, are all pre-1500. The middle quartet, also mis-ordered, are all from the sixteenth century. The rightmost quartet are scrambled seventeenth century corpora, aside from the intruding “1563”. Judged by these results, Holmstedt’s stance is reasonable.

COMPARISON OF SINGLE-FEATURE AND MULTIPLE-FEATURE SEQUENCES

Gauging the accuracy of inferred sequences

There is a standard measure of the distance of one ordered provisional list of entities from the true order, the *swap distance*.[7] In a *swap edit*, two mis-ordered adjacent items are swapped to put them in the correct order, and the distance is incremented by one. Consider the ordering resulting from the single-feature sequencing depicted in Figure 5 (omitting "1395" so the analysis is congruent with that previously obtained using five features):

One-feature	1450	1488	1413	1513	1530	1588	1543	1638	1563	1675	1613

The true order is:

True sequence	1413	1450	1488	1513	1530	1543	1563	1588	1613	1638	1675

Readers may verify that the minimal number of swaps necessary to convert the upper sequence into the lower is seven, the swap distance from the one-feature-based sequence to the true sequence.[8]

The sequencing result for the ME data found by Forbes (2016:920, Figure 16) using five-feature seriation differs by only one swap edit from the true sequence. This result is markedly superior to the one-feature result obtained above. This comes as no surprise. The general superiority of multiple-feature analysis over single-feature analysis is a truism of pattern recognition theory.

---

[7] The *swap edit distance* is also known as the *Kendall τ distance* or the *bubble-sort distance.*

[8] *Technical note*: The swap distance may be computed using AllSeqDists in the RMallow library of R.

BIBLIOGRAPHY

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

Forbes, A D 2016. The diachrony debate: a tutorial on methods, *JSem* 25/2:881–926.

_____ 2017. On dating Biblical Hebrew texts: sources of uncertainty/analytical options, in Li and Dyer 2017:297–330.

Holmstedt, R 2012. Historical linguistics and Biblical Hebrew, in Miller-Naudé and Zevit 2012:97–124.

_____ 2016. *The relative clause in Biblical Hebrew.* Winona Lake, IN: Eisenbrauns.

Li, T and Dyer, K (eds) 2017. *From ancient manuscripts to modern dictionaries*. Piscataway, NJ: Gorgias.

Miller-Naudé, C and Zevit, Z (eds) 2012. *Diachrony in Biblical Hebrew*. Winona Lake, IN: Eisenbrauns.

Rezetko, R and Young, I 2014. *Historical linguistics and Biblical Hebrew: steps toward an integrated approach.* Atlanta: SBL.