Jacob Thaisen

HOW TO MEASURE ORTHOGRAPHIC SIMILIARITYAMONG MIDDLE ENGLISH MANUSCRIPTS?

A traditional way of assessing the linguistic similarity among ME primary sources is to profile scribal usages by selective questionnaire designed to elicit diagnostic forms, especially dialectally diagnostic forms. Often phonological (<nat:not> NOT), morphological (<-th:-s> 3RD PRES. IND. SING. SUFFIX) and lexical (DARK:MURKY) variants are considered together. Angus McIntosh (1974; 1975) distinguishes between graphemic contrasts with and without phonological significance (<nat>:<not> NOT and <it:itt> IT, respectively) but has no category for phonological contrasts that have no graphemic significance. The medieval scribes themselves may have followed Donatus in thinking of a *littera* as having a name, a sound-value, and a graphic shape as distinct properties. In this classification scheme, two shapes may share a sound-value (<y:þ>) or two sound-values a shape (<h> in <thauh> THOUGH:<his> HIS).

My own work has in recent years come increasingly to take the graphic shape as its starting point and so to play down semantics and phonology. Such an approach allows for cases where a scribe inserts, say, <th> in place of an exemplar <þ> simply because his orthography did not actively include <þ>. In such a case it seems doubtful whether the scribe's practice can be described as "translation" even if the exemplar form and the copy form are unidentical. Rather, the scribe is adapting what is in his exemplar to his own *littera* system as he copies. This line of reasoning may be applied to pairs such as <eeC:eCe> or <sh-:-sch-> too, with the implication that the McIntosh distinction between "translation" and "transcription" as scribal copying modes is itself dynamic.

The most fundamental question in these classification efforts is what constitutes a variant. The increasing availability of diplomatic transcripts of ME manuscript texts opens up new, quantitative avenues that prompt a reconsideration of this question. One of these avenues is to adopt a graphic shape-centered approach by language modelling the transcripts probabilistically through building context-sensitive n-gram models of grapheme combinations.

Language modelling is a highly pragmatic approach that is widely applied in automated speech recognition, statistical machine translation and other natural language processing applications. The 3-grams that can be generated from "Canterbury Tales" are "Can", "ant", "nte", "ter"... "y T"... "les". Training texts are used to identify what 1-, 2- and 3-grams occur in them and assign frequencies to each. The resulting tables constitute a 3-gram language model of those texts, and the agreement between this model and a test text can be calculated as a measure known as entropy. It is possible to reduce the role of token frequency in favour of type frequency so as to address lexical difference, which is especially relevant when the training corpus is small, as is the case with my research.

I have applied this approach to all 58 fifteenth-century copies of Chaucer's Wife of Bath's Prologue and Miller's Tale, once with the Wife copies as model to be compared with the Miller copies as test texts and once with the opposite settings. The scores contained in the two resulting 58x58 tables were then normalised and clustered. The results confirm the existence of some manuscript pairings that are long established from palaeographic and textual work, such as Hengwrt and Ellesmere together as a subgroup of a larger cluster that also includes Christ Church and Additional 35,286.

Encouraged by these results I venture in the conclusion to this paper that probabilistic language modelling of manuscript texts can be employed, notably without much effort, by editors, palaeographers, textual scholars and historical linguists alike as a powerful tool in, for example, tracking changes of scribe or exemplar.