Method as tautology in the digital humanities

David-Antoine Williams
English Department, St Jerome's University in the University of Waterloo, Canada

Abstract

This article explores a concept of *method* in computer-assisted literary criticism, using a current digital humanities project as a case study. The project is investigating aspects of *intertextuality* between English poetry and the *Oxford English Dictionary*, second edition (OED2). In the course of adopting, applying, and adapting methods to guide computer-assisted comparisons between OED2 and poetry corpora, questions have arisen about the desired relations among research question, method, result, and outcome. Arguing that additional deliberation on what method means to us is now both appropriate and essential to the maturing discipline of digital humanities, in this article I discuss what digital methods have shown us about OED2 and Geoffrey Hill's notoriously intertextual long poem *The Triumph of Love* (1998), both as an example and an illustration of one way of reflecting on these questions: a concept of digital *method* as *tautology*.

Correspondence:
David-Antoine Williams,
St Jerome's University,
290 Westmount Rd North,
Waterloo, ON,
Canada N2L 3G3.
Email:
david.williams@uwaterloo.ca

1 Introduction

The theme of the Digital Humanities 2012 conference was 'Digital Diversity: Cultures, Languages, and Methods', reflecting an essential concern with the methods and methodologies of this expanding discipline.[1] In the modern academy, 'discipline' denotes a 'department of learning or knowledge', as the *Oxford English Dictionary*, second edition (OED2) glosses it, yet there is an older sense of the term, embedded in the ancient relation between 'discipline' and 'doctrine', the former once having described the methodical practice or training of the disciple or student, the latter the abstract theory of the doctor, or master (OED2). The attention given in the digital humanities to creating and improving various and diverse digital methods can be seen in one way to promote multi- or interdisciplinarity, by developing a commons of tools and techniques available to researchers in various fields. At the same time, however, the 'build it and see' approach, useful as it can be, avoids difficult and contentious questions surrounding disciplinarity and interdisciplinarity. One such difficulty is that, in taking methods for a department of knowledge in and of itself, practitioners in the digital humanities risk mistaking means for ends, approaches to questions for answers, ways of acquiring knowledge for knowledge itself. And, if our concern does become essentially one of practice, we must then ask at what point diverse and disparate digital methods (promoted by the phrase 'digital diversity') cease to belong to a coherent or cohering discipline, however interdisciplinary it is conceived as being. It may even be that the focus on building has ironically turned what Willard McCarty criticizes as the 'knowledge jukebox' (McCarty, 2005, pp. 6, 27) in upon itself, resulting in methods being built for their own sake, or for building more methods, without clear correspondence to existing or envisioned humanistic

doi:10.1093/llc/fqt068 Advance Access published on 30 November 2013

research questions. At the foundation of these broad disciplinary issues are the relationships created among research community (or communities), researcher, research method, and research subject. The discipline's pauses at self-reflection and self-theorization have rightly focused on describing or prescribing the nature of these relations in their different aspects (e.g. McCarty, 2002, 2005; Benyon *et al.*, 2006; Drucker, 2012; Edwards, 2012; Quamen, 2012; Ramsay and Rockwell, 2012).

In developing an idea of *method* in the digital humanities, here I must limit my discussion to the intersection of the digital humanities and the discipline of literary studies—even more specifically to the subdiscipline of literary criticism—though I hope readers in other fields will see analogies, *mutatis mutandis*, to their own disciplinary requirements and concerns.[2] One reason for this circumscription is prudence: having received as much disciplinary training in literary studies as the academy can provide, I am credentialed (a 'master', 'doctor', and 'professor', by the standards and nomenclature of the academy) to reflect critically on its epistemological orientation—its assumptions and objectives, and the methodology that leads from one to the other—including how this might interact with digital methodology and methods. To take as an example the most recent work (at the time of writing) involving literature to be published in this organ, by the same limited authority I can also give the view (however arguable it may be) that 'Ranking contemporary American poems' (Dalvean, 2015) is misinformed (or not informed) about the disciplinary concerns of poetry criticism. The ostensible subject of the research (contemporary American poems) is a mere test case for a digital method. If it is of any use at all (here I am not qualified to say), this work will be of use to researchers in machine learning and automatic classification, not in literature.

Related to this view, limiting my discussion to digital methods in literary criticism also foregrounds a particular set of epistemological principles or assumptions with which to contemplate the researcher–method–subject relation. Despite the post-Romantic project inaugurated by Matthew Arnold to establish a positivist criticism which would 'see the object as in itself it really is' (Arnold, 1864, p. 5) by adopting the epistemological claims of the natural sciences, the knowledge generated by criticism—as early as the New Criticism, and more recently and all the more radically in the work of post-structuralism—has come to resemble much more the kind of knowledge generated by creative writing than that generated by scientific method.[3] Thus Arnold's description of literary creation now seems equally applicable to criticism: it is the work of 'synthesis and exposition, not of analysis and discovery' (Arnold, 1864, p. 5). That is, literary criticism relies partially but crucially on acts of reading, which partner the critic and the text in a mutual creation of understanding. And, just as a poem on love is a highly contingent and provisional kind of knowledge about love, so a critical reading of a poem on love is a contingent and provisional kind of knowledge about the poem.

To help think about the role of digital methods within this particular relation of researcher and research subject, I take for a case study an ongoing digital humanities project at St Jerome's University in the University of Waterloo (Canada), which is primarily occupied with the detection of literary 'intertextuality'—a complex topic to which this journal has devoted a large number of pages in recent years (Trillini and Quassdorf, 2010; Forstall *et al.*, 2011; Kane and Tompa, 2011; Ben-Porat, 2012 and Coffee *et al.*, 2013 to name just a few of the most recent).[4] As with many projects in literary computing, this one came about when a researcher and a digital resource came into contact. In this case, the resource was the source data of OED2 as it was first digitized by computer scientists at the University of Waterloo in the late 1980s. The data itself had been available to researchers at the university for over 20 years, and had been the subject of several publications in the field of computer science (e.g. Berg, 1989; Townsend, 1989; Raymond, 1990; Raymond *et al.*, 1993). In order for this resource to be of value in the field of literary criticism, a literary-critical research question was required.

As it happened, this researcher had been pursuing just such a question through traditional means for several years. It was, put simply and broadly: ‘how has modern poetry been influenced by the *Oxford English Dictionary*?’ Addressing this question digitally required digital methods to be adopted, adapted, and invented. But in the execution, essential questions also arose regarding the relations among researcher, research question, method, result, and outcome. I discuss these issues in sections 4, 5 and 6, below. First I describe the electronic resource that gave rise to them, and how that resource has been exploited.

2 A Prototypical Digital Humanities Resource: ‘Crowdsourced’, ‘Intertextual’, ‘Hypertextual’

The *Oxford English Dictionary*, in its various editions, is widely considered to be the English language dictionary of record. The core of the dictionary (in the second edition) is its 2.38 million quotations, illustrating 810,456 definitions in 291,592 entries.[5]

Although the lexicographical work of compiling the dictionary took place mostly in Oxford, much of the source material for the first edition was gathered in reading rooms and private studies across the English-speaking world. Even before James Murray was taken on by Oxford University Press as editor in 1879, there had been over 100 volunteer readers (Brewer, 2008b), whose job it was to collect evidence for word usage in the form of quotations from novels, plays, poems, treatises, newspapers, and so on. Murray expanded this program greatly, publishing three *Appeals to the English-Speaking and English-Reading Public in Great Britain, America and the Colonies*. By 1884, Murray had received quotation slips from 759 volunteer members of the public (Murray, 1884). When the first volume was published in 1888, Murray acknowledged 276 of the most industrious individuals by name. Together they had sent in over 1.5 million quotation slips (Murray, 1888), almost a third of the total that Oxford's editors would consult. One man, Thomas Austin, personally contributed 165,000 quotation slips.

Not all contributors were the age's super-users, and each had somewhat different motivations and commitments. The author and poet Thomas Hardy sent in several slips recording dialect and regional usages of Wessex (Taylor, 1993, p.117) (Hardy himself is quoted 1,416 times in OED2). From time to time in the 1970s, W. H. Auden would knock up his neighbour in Christ Church, Oxford, OED general editor R. W. Burchfield, to insist that he record some word that Auden had himself just invented (Brewer, 2008a, pp. 194–5) (Auden appears 773 times). Burchfield looked fondly on idiosyncratic poetical inventions and combinations, describing them as ‘golden specks in the whole work’ (Burchfield, 1989, p. 12).

The critical point about the diverse, often idiosyncratic sourcing and selection of evidence quotations in OED2 is that the dictionary was conceived ‘on historical principles’, meaning that the quotations it cites are not merely illustrative, as they had been in Johnson (1755), though they are also that. More importantly, the 5 million source quotations (more than twice the number reprinted in the dictionary) formed the body of linguistic evidence from which lexicographers working in Oxford would deduce senses and write definitions.

So, in 1928 the world received,[6] in prototype, its first ‘crowdsourced’, ‘intertextual’, and ‘hypertextual’ work, decades before the oldest of these words would be attested in it.[7] When OED2 appeared on CD-ROM in 1992, the embedded potential of hypertextuality began to be realized digitally, with 580,632 cross-references to other sections marked up for click-through. Search and cross-referencing functionality would improve with successive digital versions, though some changes removed functionality that had existed before. In late 2012, over 130 years after Murray's first *Appeal*, Oxford University Press adapted the original practice to the language and methods of crowdsourcing, launching a new series of online *OED Appeals*, ‘where *OED* editors ask for your help in uncovering the history of particular words and phrases’ (OED.com, 2012).

The digitized OED changed completely and forever both how the work was used and how scholars came to think of it. The CD-ROM editions, and later *OED Online*, permitted quick analyses of the dictionary's contents, allowing users search for text in the body of the entries (as opposed merely to the headwords) or to search for words in use within a specific range of dates, or to count up things like how many times an author or work appeared in the quotations. All subsequent scholarship on the OED—as well as a vast amount of literary scholarship—has relied to some degree on a rich digital interaction with its data and metadata.

3 Intertextuality

In addition to a lexicon of words, definitions, and etymologies, therefore, OED2 is a book of organized quotations, drawn from tens of thousands of printed works, many of them literary.[8] The research potential of this was recognized even before digitization (Schäfer, 1980), and since has been exploited by a number of scholars in linguistics (Crystal, 2000), lexicography (Brewer, 2008a; Goodland, 2011; McConchie, 2012), and literature (Taylor, 1993). Among the many things that can be quantified by counting or sampling quotations, substantial attention has been given to nineteenth century reading habits and the social assumptions and prejudices that underlie them (such as an overrepresentation of nineteenth century texts, the importance of Shakespeare, and the underrepresentation of women authors) (Brewer, 2008a, pp. 122–30, 184–90; McConchie, 2012). All such investigations were limited by incomplete access to the data as well as by changing interfaces, which were sometimes accompanied by reductions in functionality. In principle, using the raw OED2 file, all of these studies can be redone with accuracy equal to that of OED2 itself.

Even richer interactions are possible. In addition to clarifying and extending questions of how culture shaped the dictionary, with the raw data we can now begin to document another kind of intertextuality: the OED's shaping of literature. As the English dictionary of record, it is a book for which the basic precondition of an intertextual relationship can be answered in the affirmative with a degree of confidence: we may assume at a minimum that an author writing after 1928 is aware of the OED, and further that he or she is likely to have some familiarity with it. Yet considering the dictionary itself as a potential intertext complicates somewhat the idea of intertextuality, since there are several ways in which information recorded in the OED can find itself activated in a poem or other literary text: (i) the poet may consult OED to verify or investigate a word's meanings, etymologies, and historical uses; (ii) the poet may consult the OED as a source in itself, thinking critically or poetically about what information it presents, and how; or (iii) because OED is already an intertextual source, the poet may draw identical or similar information from another source—general knowledge, perhaps, or an etymological dictionary, or the original source text of an evidence quotation. This final relation can be seen as a kind of intertextuality, though perhaps a reduced kind, in that it brings into a literary text information about the life of one or several of its own words, information that may appear in any number of works, but which also appears in OED.

4 Case Study: Geoffrey Hill

Representative of a pervasive view of Geoffrey Hill's poetry is the statement that it is 'notably armoured in learning, and often intensely difficult because of its dependence on etymology and allusion' (Mackinnon, 1997, p. 23). One of Hill's favoured sources is OED2. He has described himself 'brooding' over its pages, saying that 'Most of what one wants to know, including much that it hurts to know, about the English language is held within these twenty volumes' (Hill, 2008, p. 279). Hill's work therefore presents an ideal case study for computer-assisted detection of intertextuality with the OED.

For this study, two search and comparison programs were written in the Python programming language, using techniques adapted to the nature of the text contained in two distinct sections of OED2 entries:

(A) **Etymology Match**. The first program reads and filters a target text, retrieving the OED2 etymology section (also filtered) of every word in the target text (lookup word), or its lemma (determined using the Python 'NLTK' module) (Bird *et al.*, 2009) or word stem (using the Python 'stem' module's Porter2 algorithm). It then looks within a user-specified span of words in the poem for any words (or their lemmas, etc.), which also appear in the etymology section in question (match words). If a match is found, the program scores it based statistical measures of the likelihood of the match occurring in various genres of corpora (in this case, a poetry corpus, the British National Corpus, and a corpus of all OED2 quotations). *Prima facie*, less likely matches are expected to be more significant.

(B) **Quotation Match**. The second program reads and filters a target text, generating lists of word-level *n*-grams (in this case, bigrams, trigrams, and four-grams). Going by overlapping subsections of a specified size, for each word, its lemma or stem (lookup word) in the target text, the program then generates *n*-gram lists for all filtered OED2 quotations under that headword. Tokens in the two *n*-gram lists are then measured against each other for similarity (0–1.0, in which 1.0 is an identical match) using the Python 'difflib' module, and matches above a user-specified threshold reported, along with statistical scores similar to the ones described in (A).

Both programs were run using Hill's notoriously allusive collection *The Triumph of Love* (Hill, 1998), a single long poem of 150 sections, containing about 10,850 words, as the target text. Results were then ranked according to their various metrics and inspected manually.

Using section-specific stop-word lists developed for this search, the Etymology Match program found 439 lookup words in the poem within 10 words of a match word appearing in the lookup word's OED2 etymology. On inspection, matches fell into one of three categories: (1) unassociated, where the match is due to a merely functional word in the lookup word's etymology and/or is an artefact of the stemming algorithms (e.g. 'both-follows', 'honour-write', 'help-rooted'); (2) trivially associated, where a semantic or collocational relationship exists but not a significant etymological relationship (e.g. 'tit-tat', 'year-day', 'indigo-dye'); and (3) etymologically related ('elm-ulmus', 'cordial-heart', 'luminaries-lumen', 'conversion-turned').

Twenty-one pairs were determined to belong to the third category, and these were then inspected in context. Six sections of the poem were deemed to be creating etymological tropes, often using more than one of the matched pairs detected by the program. Three examples can illustrate different kinds of poetic techniques behind what the program detected:

A. Matches: 'cordial-heart'; 'courage-heart'; 'courage-heart'

> Southwell, addressing the cordial
> cordially: 'it does my heart good'.
> Fifty years without limbs, or in an iron
> lung, is that possible? I lose
> courage but courage is not lost.
> (Hill, 1998, p. 22)

Here the lookup words 'cordial' and 'courage' both match with 'heart', the English word for their Latin etymon *cor*. The first match is an etymological play on words, where 'the cordial' refers at the same time to 'matters of the heart', something (such as a beverage) that 'invigorates the heart', and also those who are 'warm or friendly'. In context it may be imagining the courage of Southwell's address to the crowd at his execution and subsequent disemboweling. Though 'cordial' is partially marked as an etymological pun in the text, 'courage' is not, forming instead a kind of etymological echo behind the audible echo of 'cor' down through the passage.

B. Match: 'casus-fall'

> The Florentine
> academies conjoined
> grammar and the Fall, made a case of *casus*.
> (Hill, 1998, p. 75)

Here an old association between grammatical case and *casus hominis*, the Fall of Man, is invoked. Though the etymological relation between 'case' and *casus* is flagged in the text, the meaning of Latin *casus* ('fall'), and therefore the full relation between 'case' and 'fall', may remain obscured.

C. Matches: 'converte-naturally'; 'converte-turn'; 'convert-naturally'; 'convert-turn'; 'conversion-turned'

> *Oculos tuos ad nos converte*: convert
> your eyes, *Vergine bella*: you gave us
> a bit of a turn there. Not unnaturally —
> but not naturally, either —
> [. . .]
> Since when has our ultimate reprobation
> turned (*oculos tuos ad nos con-*
> *verte*) on the conversion or
> reconversion of brain chemicals
> (Hill, 1998, p. 56)

Here the etymological relation among Latin *converte*, English 'convert', 'conversion', and 'turn' is deliberately signalled, turned upon, and returned to in the poem. Yet as the cluster of repeated matches in the program's output draws attention to this area of the poem, the appearance of 'naturally' as a match word reminds the critic that to 'convert' is, according to OED2, to '*turn* in character or *nature*', that 'nature' is built into the etymological substrate of the word.

While the Etymology Match program points to areas in the poem where information that OED2 holds about the English language is activated, it is difficult to imagine many likely scenarios in which it will demonstrate any direct influence of the dictionary on the poet. Indeed the very fact that we can recognize etymological tropes at all indicates that they may also draw to some degree on common knowledge and intuitions. Those matches that appear least likely on inspection, therefore, if they turn out to be significant, may be the best candidates for claiming influence.

The same basic problem of intertextuality applies to the second program, which compares a target text with OED2's evidence quotations. Two examples from the program's results can illustrate the difference between the strong claim of direct influence and a reduced claim of broad cultural intertextuality. The program found seventeen trigrams in the poem matching (with similarity >0.900) trigrams in the OED2 quotation sections for nearby lookup words in the poem. Several, on inspection, were found to be unassociated (e.g. 'free expression poorly' was matched to 'free expression play' s.v. 'free', with similarity of 0.905), or trivially associated because of a semantic or collocational relationship (e.g. the idiom 'round half dozen' matches exactly s.v. 'round'). Of the seventeen, five were deemed to be significant:

Lookup word	Poem trigram	OED trigram	Similarity
striker	divine striker senses	diuine striker sences	0.905
mark	removeth neighbour mark	remoueth neghbours mark	0.913
king	Hudson Railway King	Hudson railway king	0.947
mercury	Salt sulphur mercury	salt sulphure mercury	0.975
tautology	tautology vain repetition	tautology vain repetition	1.000

To take the first four examples in context, the poem itself marks their intertextuality in different ways: the first ('divine striker upon the senses' in the poem) is attributed to Sidney and placed in quotation marks; the second ('*Cursed be he who removeth his neighbour's mark*') is italicized within a discussion of Tyndale (though OED2 quotes the Coverdale version); the third ('[Hudson the Railway King –ED]') appears as it were an editorial note; the fourth ('Salt, sulphur, | mercury') comes just after the mention of a seventeenth century chemistry treatise. That these trigrams appear in OED2 is no sign that Hill found those phrases there, only that both Hill and the contributors to OED2 have read the same or similar works and deemed them significant, though likely not always for the same reasons or purposes. This knowledge is not trivial to the critic, since it adds a dimension to the intellectual history of the phrase alluded to.

The final example is a far more complex instance of intertextuality. The section of the poem where it appears is as follows:

> Estrangement itself
> is strange, though less so than the metaphysics
> of tautology, which is at once *vain*
> *repetition* and *the logic of the world*
> [Wittgenstein].

A few lines later the poem returns to these ideas:

> Tautology
> for Wittgenstein, manifests the nature of
> unconditional truth. Mysticism is not

> affects but grammar. There is nothing
> mysterious in grammar; it constitutes
> its own mystery, its *practicum*. Though certain
> neologisms – Coleridge's 'tautegorical'
> for example – clown out along the edge
> (Hill, 1998, pp. 66–7)

As in the instances discussed above, here *vain repetition* and *logic of the world* are being marked in the text as intertextual references, set in italics, with the name Wittgenstein appearing both as an editorial incursion and a discursive subject.

A reader might fairly assume from the context that Wittgenstein is the source of both italicized quotations, but this would be wrong, for there are at least two unacknowledged sources. The phrase 'the logic of the world' does indeed come from the *Tractatus Logico-Philosophicus* (Wittgenstein, 1922, s.6.22). But 'vain repetition' is not to be found in that work. It comes from a considerably more obscure text by the seventeenth-century English minister and politician William Gouge, his *Commentary on the Whole Epistle to the Hebrews* (Gouge, 1655): 'there is no tautology, no vain repetition of one and the same thing therein'. And this is not the only uncited source in the passage: as the program has detected, these words appear in the quotations s.v. 'tautology' in OED2. Also in the OED2 evidence quotations we find, not Wittgenstein's 'the logic of the world', but rather 'The tautology..is unconditionally true' (Wittgenstein, 1922, s.4.461), which makes a partially obscured allusive appearance in the second part of the excerpt above, when Hill writes, 'Tautology | for Wittgenstein, manifests the nature of | unconditional truth'. This allusion was also detected by the program at the bigram level, matching 'unconditional truth' to 'unconditionally true' s.v. 'tautology', with a similarity of 0.872.

So in these lines, setting aside the references to Augustine and Pascal, there are no fewer than four sources mixing intertextually: the *Tractatus*, the *Tractatus* and Gouge's *Commentary* as quoted in OED2, and OED2 itself—certainly the rest of that entry, including the definitions, and perhaps related or nearby entries as well ('tautegorical', almost certainly, which OED2 credits Coleridge with inventing). It is also possible that a fifth work, Gouge's original text, has been brought to bear on the passage, prompted by the appearance of the quotation in OED2. All this is potentially valuable knowledge for the critic, especially if the research question investigates the role of dictionaries in the making of poems. And if such evidence of direct influence of a dictionary on a poem can occur here, the critic must wonder where else it has occurred, and how much will have to be read, looked-up, and cross-referenced, before another example is found.

5 Method

The outcomes of these two programs may seem obvious in retrospect, once our eyes have been turned in their direction. Ideally, many *will* seem obvious, or inevitable—something one can hardly believe one could miss. But of course they are not always obvious on first or even on repeated close readings of the text, especially where the density of allusion is such that any line may point in one or more of several different directions, or none—where there is no way of knowing when a line's allusive potentialities have been exhausted. The manual investigation of etymologies and allusions, even on a small amount of text, is time-consuming and may often be fruitless, even if the disciplinary training and linguistic intuitions of the researcher reduce vastly the number of searches that will be performed. By increasing vastly the number of searches, identifying 426 etymology matches and seventeen quotation trigram matches among the 10,850 words, and by providing statistical guides to their likely significance, the method here decontextualizes the linguistic relation, asking the critic to reassess what may have been passed over. And the field of English literature is large: in the case of a corpus of texts, these methods allow selective perusal of what would otherwise be simply too much text to read attentively.

A process of feedback between the critic's requirements and the program's provisional results will improve the method by iteration: in this case refining the algorithm, text filtration techniques, and corpus-based metrics will reduce noise in the results, drawing the critic's attention more closely to the areas that resemble those that have been deemed significant in the past. Alternatively, the failure to identify an expected match may lead the researcher to reconsider more fundamental aspects of the method, so that it correspond more closely to the researcher's understanding of the research activity. Or, the method may capture a class of instances that meet the criteria described by the researcher, but which the researcher rejects as irrelevant to the desired outcome, forcing him or her first to realize and analyse the pertinent distinction (what is the likeness, and what is the difference, between result and desideratum?), and then encode it into the method. Lastly, the method may even—if rarely—return results that force the researcher to turn back upon the original research question driving the inquiry.

As essential to achieving research outcomes as this iterative process is, the method's provisional results are not research outcomes per se. The improvement of the method is para-investigational, in the sense that it has for its primary objective a narrowing of the gap in correspondence between research question and the method's desired outcomes. Willard McCarty has argued that 'Humanities computing lives and deals in that gap' (McCarty, 2002, p. 104), and it is true that the development and refinement of the method, along the lines of McCarty's idea of *modelling*, is an intellectual activity that is 'neither solely computational nor autonomously human but a combination or interaction of both—a thinking with, and against, the computer' (McCarty, 2002, p. 104). In each iteration, a program's results will show both things of relevance to the question, and also the ways in which the program has failed to address the research question. Improving the method's results may require technical improvements or corrections, but it is at least as likely that the researcher will be lead to reflect more fundamentally on the way the intended outcome has been described in the grammar of the method, how he or she has abstracted and translated this intended outcome, or modelled the thing being investigated.

Yet McCarty's claim that 'properly speaking a model teaches us when it fails to correspond to what we expect of it' (McCarty, 2002, p. 105) is only true in this limited (though not inconsequential) arena.[9] As he says, a successful model 'might be of interest as a process that generates useful results for some other purpose' than the study of models and modelling (McCarty, 2002, p. 105). But we might change 'might be' for 'ought to be'—for the humanities researcher, models and methods themselves are not always or even usually the primary purpose of the enquiry. However much a failure to deliver expected results will teach us about the grammar of the programming language, and however deep the reflection it will provoke on the embedded grammars of our own thought—that is, however much it requires the researcher to define and describe his or her own research intentions in another grammar—by definition the failed method does not teach, or teach enough, *about the research question* driving the activity. The 'provisional, contingent nature of a continuing activity' applies in this limited arena only while a result sufficient to the intended outcome remains elusive. Of course, it may be the case (e.g. for researchers in the philosophy, history, and sociology of information, education, or science) that the research question is precisely about the integration of digital methods with the human sciences, the model considered *as model*, a topic for research and reflection in itself (McCarty, 2002, p. 105)—in which case all models, all methods are always teaching in one way or another. Researchers based in other humanities fields will want the method eventually to contribute to their disciplinary work, to produce an outcome corresponding in some useful way to what was expected, or desired, when the method was first provisionally and contingently conceived. This is 'delivery' in the sense rejected by McCarty, but it is not '*mere* "delivery"' (McCarty, 2005, p. 6, my emphasis). And while delivery does not necessarily put an end to the work of developing and refining a method, it should be considered the primary end of method development.

It is from the model-as-model (or method-as-method) perspective that Benyon *et al.* approach the question of how best to 'integrate automated processing with human thinking and acting' (Benyon *et al.*, 2006, p.142). Describing a 'servant or partner' dichotomy in how the relation between digital methods and human researchers can be conceived, they sensibly argue for the latter (leaving unexamined, however, the question of what else a master might gain from such a relationship, in addition to services rendered, and not seeing, apparently, the possibility that a method may itself act as 'master'—see again Dalvean, 2015). But 'partner', though preferable, is not adequate, unless the method is itself the subject of research, as a poem and a critic can be said to act together in the mutual creation of understanding.

Beyond this limited arena it may be better to analogize McCarty's 'a model teaches us...' to how professors will sometimes piously assert that 'my students teach *me*'. No one who has taught would dispute that in the activity of teaching, a teacher learns many things—about the interests, desires, capabilities, and limitations of students, for instance, and, importantly, about what pedagogical methods work best to explain and instill disciplinary knowledge and methodology. More than this, insights into disciplinary concerns may emerge during class preparation or in the classroom—may even be suggested to the teacher by a student's intervention. But this is not to be confused with the student schooling the master in the discipline. With the important exception of the researcher in pedagogy or education science, for whom the classroom is always potentially a place of experimentation or observation, teaching is a para-disciplinary activity. Ideally it will lead a teacher to reflect on and perhaps improve his or her disciplinary knowledge and training, but it is not disciplinary training or practice per se. Similarly when a researcher describes an algorithm in a programming language, or builds a user-defined variable into the algorithm, or chooses to apply a particular statistical measure, he or she is building a method that accommodates extra-disciplinary techniques to disciplinary knowledge and training, adapting those techniques to disciplinary requirements while also learning how best to communicate those requirements in that foreign grammar. Not servant–master, therefore, nor partner–partner, but student–master, or disciple–doctor.

6 Tautology

Like all metaphors, the researcher:method::teacher: student analogy must at some point fail. However, as with all metaphors, the point of failure may turn out to be as instructive as the span of correspondence. While I believe understanding the relation in these terms gives a strong metaphor for the disciplinary reflux or feedback of a para-disciplinary activity, the analogy is less robust on the other side: there are important ways, germane to this discussion, in which a method is wholly unlike a student. One of these—the most obvious perhaps—is that the student is not created by the teacher in order to further a research enquiry or produce research outcomes; a method is. And, though one might say that in the abstract a teacher sets the terms of a student's learning, in actuality teachers hope and expect that students will rearrange and recombine information from various areas of their lives and learning. The failure of a student to integrate disciplinary knowledge in the expected way might be the result of a limitation of his or of his teacher's capacities, but it might also be the result of human genius, a cognitive leap of a kind as yet unobserved in computational processes. Another metaphor is therefore to be desired, which would better figure the epistemological horizons of the method and the relation of these to the researcher's own epistemological horizons. To develop this metaphor, I return to an outcome of the digital method applied to Geoffrey Hill's long poem: the passage detected by the Quotation Match program in which Hill, using OED2 quotation evidence, extends and broadens the poem's understanding of 'tautology'.

That disciplinary reflection and development could be generated by computing activity did not occur to Hill when he wrote, in a review of OED2 for the *Times Literary Supplement*, that 'the computer is now operating in the interests of cohesion.... If there had been an original bias or imperception ... I would not now expect it to be reconsidered' in the electronic version (Hill, 2008, p. 278). Driving home the point, he asked his reader to consider, 'In what sense or senses is the computer acquainted with original sin?' (Hill, 2008, p. 279). The printed OED2 is of course acquainted, and may acquaint us, with 'original sin' in one sense at least, reflected in its definition (s.v. 'original', 1b: 'the innate depravity, corruption, or evil tendency of man's nature') and the nine quotations that illustrate it. The electronic version may familiarize us with additional senses, however: the occurrence of 'original sin' within the definitions of seven other headwords, and in thirty-five other quotations illustrating senses of headwords from 'birth' to 'yatter'. But to insist on this is to evade the point Hill is making, that the computer itself is not familiar with any sense of 'original sin' at all, beyond those that have been supplied to it at its own moment of origin. In reacquainting us in a matter of seconds with several senses which would have taken years for a human to compile from the dictionary (and which *did* take hundreds of person-years to compile *into* the dictionary), the computer gives the impression that it is making a discovery, yet it is only reminding us of something already known—texts that have been read, selected, compiled, abstracted, and indexed by the reading public and lexicographers. There is nothing that the computer will do to think through, as we must, the implications of Hill's sense or senses of 'original sin'. Nothing in the computer will think, as Hill does in his poem, about the sense or senses of 'mystery', 'grammar', and 'tautology', and how these may relate to one another, and how the poem relates them together.

But thinking about these terms—along with Hill, but in some ways against him too—may enrich our understanding of the method that pointed us to them in the first place. One common metaphor for a digital model or method has been, for obvious reasons, that of a 'grammar', a comparison that has been extended in several ways, including towards Chomskian transformational grammar (McCarty, 2005, pp. 24, 55). It may help to understand the limitations and capacities of the computer's grammars to start thinking of a *method* in the digital humanities as a *tautology*—not in the pejorative sense applied to the rhetorical fallacy, something necessarily and so trivially true, but in the sense Hill's poem develops, a 'grammar' that 'constitutes its own mystery, its *practicum*'. If the 'world' of a computer program is the inputs it receives, the program itself is the 'logic of the world', a *tautology* in the Wittgensteinian definition pondered by Hill. As a *tautology*, a *method* will arrange and combine the elements of the world it is given according to its own prescribed logic, a rearrangement that may correspond more or less, depending on its grammar, to the logic and logics of our world.

Thus, for example, a word cloud created from a text will correspond differently to our habitual perception of that text than a table of contents will, a graph differently than an index, a ranked table of etymologically related pairs differently than the same pairs of words occurring in place in the linear text—though all are reconfigurations of the original textual matter. The value of the rearrangement will be in how the critic perceives the rearranged world in relation to his or her preconceived world (be this preconception naive, disciplinary, or heuristic), whether attention is drawn to a connection or disconnection in this world, which has been overlooked or underappreciated because human grammars of perception or disciplinary logics have obscured it. In other words, the value of the rearrangement is in the potential of the researcher to understand it in terms of the research question lying at the origin. It is the originating research question that relates representation to original text. Outside this basic relation, producing differences and similarities between ways of arranging and perceiving knowledge, the method is in itself meaningless. Meaning emerges from an encounter with the original research question returned to the researcher in unfamiliar guise, but depends crucially on an essential identity of the familiar and unfamiliar ways of experiencing the research object. The tautological method produces a *seeming* difference, ultimately reducible to an aspect of identity, which gives rise to understanding in the researcher's very reduction from difference to identity.

Put more simply, the method is tautological because it will not add to the world it is given, and will not alter its own grammar, its own logic, which has been described for it by a human mind. It is that human mind that must judge of the similarities and differences between grammars by 'thinking with, and against', the computer. It is, in other words, coming up against and assessing defamiliarized versions of its own thought, thinking *with* the computer, yes, but *against itself*. The real and productive questions for those contemplating the disciplinary nature of the digital humanities are therefore not to be 'with which methods should be acquaint ourselves?', nor 'what new knowledge may such and such methods acquaint us with?', but rather 'in what unfamiliar forms and grammars will we encounter our own thoughts, the more familiar we become with our digital methods?' And, 'at what point does our familiarity with digital methods itself produce habits of perception, which must themselves be broken?' And finally, related to this perhaps, 'at what point if any must we break off this acquaintance, returning (with new insight, perhaps) to the disciplinary ways of discovering and understanding that brought us to the method in the first place?' These questions are themselves unresolvable, as is humanistic inquiry—it is the asking that constitutes disciplinary reflection. As long as they hang over whatever particular research questions are being investigated, the digital humanities may continue to claim humanistic disciplinarity.

In some ways my account of method may appear to endorse a pervasive critique of digital humanities scholarship often heard from humanists outside the field, that digital approaches add nothing to our humanistic understanding of the subject matter, and may even blind us to the fine particularities that close and patient reading trains one to sense and attend to. Hill has articulated such a view: 'I think there are things built into the information culture which are destructive of the very things it seeks to gain information about' (quoted in Sperling, 2011, p. 334). He says, 'the one thing computer technology does is in fact a velocity thing—you now do in two seconds what earlier scholarship would have taken two or three years to do. A plethora of information speedily acquired is the sort of velocity that will destroy criticism, and it is a very frightening prospect' (quoted in Sperling, 2011, pp. 333–4). Yet, a concept of the digital method that concedes (even promotes) its tautological nature, in the extended and positive sense developed here, also gains a robust defence against this charge. This is because, although it cannot add to our world, the method also will not take away from it. If we understand the outcome as a rearranged and recombined knowledge—knowledge translated into another grammar, to be sure, but into a grammar described by a human and so existing ultimately in some defined relation to human grammar—then we cannot claim that it creates or destroys knowledge of our world in and of itself. The method, in other words, gives back a particular view of its world, and therefore of ours, but it will not show the critic anything the critic has not *somehow* asked to be shown, even if at times that 'somehow' will itself require (and so repay) investigation. The critic is left with the same poem in essence, the same research question in essence, and, in addition to this, the critic is left with the intellectual gains of reconciling these to the outcomes of the method.

What to do with these gains? To return to Hill's particular question—'In what sense or senses is the computer acquainted with original sin?'—one might justifiably respond that any number of computer interfaces with OED2 will remind a user of various historical senses of 'original sin'; any number of methods may help trace the intellectual lineage between others who have written on the subject and Hill. And, one might add, additional methods might remind us of Hill's related acquaintanceships with the dictionary he is discussing: his own poetic conjunction of 'grammar and the Fall', incorporating original sin into language via Latin etymology; his meditation on the 'conversion | or reconversion' of original essential nature; his deliberations on the self-mystery of tautological grammar, returning truth to its original self.

In so acquainting (or reacquainting) us with these interrelated conceptual turns and returns in the dictionary, in Hill's *oeuvre*, and in the nourishing literary and cultural history, digital methods will have equipped us to ponder Hill's rhetorical question, which in this case operates in a vaster and vastly different world than a tautological method will accommodate. This is, not least, because of the semantically shaded term 'acquainted', and its relation to 'original sin', which in Hill's deployment wavers ambiguously on the line that analytic philosophers call the 'use-mention distinction'. The grammar of the critic, developed over the history of the discipline, recognizes and accommodates the tense and productive simultaneity of multiple meaning and implicature. The digital method, as yet, does not. Rich ambiguity—simultaneous, undecideable multiple meaning—is the literary feature that stands as the challenge *par excellence* to computer modelling in literary criticism.[10]

Funding

This work was supported by the Social Sciences and Humanities Research Council of Canada [430-2011-0630]; St Jerome's University in the University of Waterloo; and the University of Waterloo.

References

Arnold, M. (1864). The Function of Criticism at the Present Time. In *Essays in Criticism*. London: Macmillan, pp. 1–41.

Ben-Porat, Z. (2012). Allusive inter-textuality in computer games. *Literary and Linguistic Computing*, **27**(3): 261–71.

Ben-Porat, Z. (1976). The poetics of literary allusion. *Poetics and Theory of Literature*, **1**: 105–28.

Benyon, M., Russ, S., and McCarty, W. (2006). Human computing—modeling with meaning. *Literary and Linguistic Computing*, **21**(2): 121–57.

Berg, D. L. (1989). *The Research Potential of the Electronic OED2 Database at the University of Waterloo: A Guide for Scholars*. Waterloo, ON: UW Centre for the New Oxford Dictionary and Text Research.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Brewer, C. (2008a). *Treasure House of the Language: The Living OED*. London: Yale University Press.

Brewer, C. (2008b). Examining the OED—Initial Practice. http://oed.hertford.ox.ac.uk/main/main/content/view/89/233/ (accessed 30 September 2012).

Burchfield, R. W. (1989). *Unlocking the English Language*. London: Faber.

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R., and Jacobson, S. L. (2013). The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, **28**(2): 221–8.

Crystal, D. (2000). Investigating Nonceness: Lexical Innovation and Lexicographical Coverage. In Boenig, R. and Davis, K. (eds), *Manuscript, Narrative and Lexicon: Essays on Literary and Cultural Transmission in Honor of Whitney F. Bolton*. Lewisburg: Bucknell University Press, pp. 218–31.

Dalvean, M. (2015). Ranking contemporary American poems. *Digital Scholarship in the Humanities*, **30**(1): 6–19.

Drucker, J. (2012). Humanistic Theory and Digital Scholarship. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. http://dhdebates.gc.cuny.edu/debates/text/34.

Edwards, C. (2012). The Digital Humanities and its Users. In Gold, M. K. (ed.), *Debates in the Digital Humanities.* http://dhdebates.gc.cuny.edu/debates/text/31.

Forstall, C. W., Jacobson, S. L., and Sheirer, W. J. (2011). Evidence of intertextuality: investigating Paul the Deacon's *Angustae Vitae*. *Literary and Linguistic Computing*, **26**(3): 285–96.

Gakis, P., Panagiotakopoulos, C., Sgarbas, K., and Tsalidis, C. (2015). Analysis of lexical ambiguity in Modern Greek using a computational lexicon. *Digital Scholarship in the Humanities*, **30**(1): 20–38.

Goodland, G. (2011). 'Strange deliveries': Contextualizing Shakespeare's first citations in the OED. In Ravassat, M. and Culpeper, J. (eds), *Stylistics and Shakespeare's Language: Transdisciplinary Approaches*. London: Continuum.

Gouge, W. (1655 [1866]). In Smith, T. (ed.), *Commentary on the Whole Epistle to the Hebrews*. Edinburgh: James Nichol.

Hill, G. (2008). *Collected Critical Writings*. Oxford: Oxford University Press.

Hill, G. (1998). *The Triumph of Love*. London: Penguin.

Irwin, W. (2001). What is an allusion?. *The Journal of Aesthetics and Art Criticism*, **59**(3): 287–97.

Kane, A. and Tompa, F. W. (2011). Janus: the intertextuality search engine for the electronic *Manipulus florum* project. *Literary and Linguistic Computing*, **26**(4): 407–15.

Kristeva, J. (1969). *Séméiotikè: Recherches Pour Une Sémanalyse*. Paris: Seuil.

Leddy, M. (1992). The limits of allusion. *The British Journal of Aesthetics*, **32**: 111–4.

Machacek, G. (2007). Allusion. *PMLA*, **122**(2): 522–36.

Mackinnon, L. (1997). The matter with England. *The Times Literary Supplement*, **23.**

McCarty, W. (2005). *Humanities Computing*. London: Palgrave Macmillan.

McCarty, W. (2002). Humanities computing: essential problems, experimental practice. *Literary and Linguistic Computing*, **17**(1): 103–25.

McConchie, R. W. (2012). 'Her word had no weight': Jane Austen as a lexical test case for the *OED*. *Dictionaries: Journal of the Dictionary Society of North America*, **33**: 113–36.

Miller, J. H. (1977). The critic as host. *Critical Inquiry*, **3**(3): 439–47.

Murray, J. A. H. (1884). The President's address for 1884. *Transactions of the Philological Society*, 501–642.

Murray, J. A. H. (1888). Appendix to Preface. In Murray, J. A. H. (ed.), *A New English Dictionary on Historical Principles*, **vol. I. (A-B)**. Oxford: Clarendon Press, pp. xv–xvi.

OED.com. (2012). About the OED appeals. http://public.oed.com/the-oed-appeals/about-the-oed-appeals/ (accessed 15 October 2012).

Quamen, H. (2012). The limits of modelling: data culture and the humanities. *Scholarly and Research Communication*, **3**(4). http://src-online.ca/index.php/src/article/viewFile/69/194.

Ricks, C. (2002). *Allusion to the Poets*. Oxford: Oxford University Press.

Ramsay, S. and Rockwell, G. (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. http://dhdebates.gc.cuny.edu/debates/text/11.

Raymond, D. R. (1990). *A Potpourri of Prototypes*. Waterloo, ON: UW Centre for the New Oxford Dictionary and Text Research.

Raymond, D. R., Tompa, F., and Wood, D. (1993). *Markup Reconsidered*. Waterloo, ON: UW Centre for the New Oxford Dictionary and Text Research.

Schäfer, J. (1980). *Documentation in the* OED: *Shakespeare and Nashe as Test Cases*. Oxford: Clarendon Press.

Sperling, M. (2011). The trouble of an index. *Essays in Criticism*, **61**(4): 325–37.

Taylor, D. (1993). *Hardy's Literary Language and Victorian Philology*. Oxford: Clarendon Press.

Townsend, G. (1989). *Citation matching in the Oxford English Dictionary*. Waterloo, ON: UW Centre for the New Oxford Dictionary and Text Research.

Trillini, R. H. and Quassdorf, S. (2010). A 'Key to all quotations'? A corpus-based parameter model of intertextuality. *Literary and Linguistic Computing*, **25**(3): 269–86.

Wellek, R. (1978). The new criticism pro and contra. *Critical Inquiry*, **4**(4): 611–24.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Ogden, C. (trans.). London: Keegan Paul.

Notes

[1] For a selection of articles from this conference, see the special issue of *Literary and Linguistic Computing* **28**(4) (December, 2013). A book of abstracts is archived here: http://www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf.

[2] The degree to which they do may be an indication (among other things) of how interdisciplinary the digital humanities are, or are prepared to be, at the present time.

[3] The classic account of New Criticism's anti-positivism is Wellek, 1978. Miller, 1977 represents one important formulation of the later radical view.

[4] Although computer-assisted work on intertextuality shares a number of stylometric and statistical techniques with attribution studies, the two differ widely in their assumptions and goals. Having applied the term more broadly than its initial description (in Kristeva, 1969) would admit, the concern of literary studies with intertextuality now involves making difficult and unsettled disciplinary distinctions among quotation, reference, reply, allusion, and influence (among other types of textual interrelation) and judging of the implications of these for criticism. See, among others, Ben-Porat, 1976; Leddy, 1992; Irwin, 2001; Ricks, 2002 and Machacek, 2007.

[5] All figures concerning OED2 have been arrived at using the OED2 (1989) data file. The data are published by Oxford University Press.

[6] OED fascicles had been published since 1884 (when the project was still known as the *New English Dictionary*), first at a rate of about one per year, and then more frequently. The final fascicle, as well as the first complete bound set of twelve volumes, was published in 1928, followed by a thirteen volume set (including a Supplement) in 1933.

[7] The OED cites T. H. Nelson as the first person to use 'hypertext' (1965) and Julia Kristeva as the coiner of 'intertextual' (1969, first used in English in 1973). 'Crowdsource', 'crowd source', which appears to have been coined in *Wired* magazine in 2006, is not recorded in the current *OED Online*, though it does appears in two other Oxford dictionaries. An editor at OUP reports that at the time of writing it is uncertain whether the word will eventually be included in the OED. As with all words, the concept necessarily precedes the coinage, let alone the first print usage.

Though the retrospective application of these terms here may appear to court anachronism, it is intended to underline a continuity between the compositional principles behind OED and its current digital uses.

[8] *OED Online* estimates the number of works quoted in the first edition to be around 4,500, a figure the Supplements would have increased considerably by the time of OED2. Counting unique works in OED2 is difficult to do automatically, since the same work can often be represented by several tokens. *Hamlet*, for instance, appears as <W>Ham.</W>, <W> Ham.</W>, <W>Ham</W>, <W>Haml.</W>, and <W>Hamlet</W>. There are 240,842 unique <W> tokens in the file, 203,948 of which occur five times or fewer.

[9] I discuss 'model' [following McCarty's definition of a 'model of' (McCarty, 2005, p. 24)] and 'method' interchangeably here, not because they are indistinguishable but because they exist in analogous relation to the researcher, and depend on the same basic questions about that relation.

[10] I do not preclude its possibility in principle, but to achieve this would require rejecting or rethinking virtually every heuristic currently employed in textual computing, which, to the degree that it has addressed ambiguity at all, so far has limited itself to the much simpler case of lexical ambiguity, and has focused entirely on methods for resolving it for purposes of natural language processing and machine learning [see, e.g. Gakis *et al.* (2013) for just the most recent example].