From token to exegesis part 1: introducing Corpus Criticism

Mark Boardman
m.boardman@hud.ac.uk
markboardman@outlook.com
+44 7780 515377
PhD Candidate and Fellow in English Language and Linguistics at The University of Huddersfield (UK)

Abstract
This paper reports on progress towards the development of a new interdisciplinary analytical method of textual study, called Corpus Criticism, that licenses reasoned subjective critical inferences along with interpretive autonomy, from computational analyses of tokenised literary corpora, combining methodology from corpus stylistics, literary criticism and natural language processing, building on and developing some elements of methodology first outlined at PALA 2019 (Boardman, 2019) in relation to my current PhD research on persona and agency in Emily Dickinson. The overall aim is to demonstrate workflows for attempting to solve the issue explored by McIntyre and Walker (2019) of how to transition from replicable corpus based analysis to meaningful qualitative judgements on aesthetic, cultural and psychological signification in textual data. Analytical and critical workflows modelled in the paper will suggest ways of linking persona and grammatical agency to prototypical and emergent evidence of a modernist consciousness (Brown, 1989; Sotirova, 2013) in Dickinson's writing.

Keywords
persona; agency; corpus; literary; criticism; stylistic; computational; NLP; grammar; Modernism

Computer do my English Lit homework

Computer scientist John B Smith's (1978) article "Computer Criticism" proposed the notion that a computer might be capable of doing literary criticism – an activity traditionally reserved for humans. As far as the electronic design and functioning of all computers is concerned, human language is processed as a quantal system. Smith points out that machines can only process signifiers and can do nothing with the Derridean (1967, 2016) notion of a signified, where 'signified' indicates humanly inferred social, psychological and cultural meaning and 'signifier' is taken to indicate a stream of machine-processible linguistic tokens. Therefore, within the limits of current design parameters, computers are handling a digitised representation of Chomskyan (1965, 2015) surface form. This process is referred to in Natural Language Processing as 'tokenisation'. (See Silberztein (2016, pp. 227-238) for an explanation of the relationship between linguistics and the storage and processing capabilities of computer software.)

Computers process: they do not analyse – a point alluded to in Harnad's commentary on Turing's seminal paper "Computing Machinery and Intelligence" (1950, 2008):

> 'We know now that what he will go on to consider is not whether or not machines can think, but whether or not machines can do what thinkers like us can do – and if so, how. Doing is performance capacity, empirically observable. Thinking is an internal state. It correlates empirically observable as neural activity (if we only knew which neural activity corresponds to thinking!) and its associated quality introspectively observable as our own mental state when we are thinking. Turing's proposal will turn out to have nothing to do with either observing neural states or introspecting mental states, but only with generating performance capacity indistinguishable from that of thinkers like us.' (p. 23)

This is an important founding principle of the science of Natural Language Processing, that machines are processors but not thinkers and are therefore incapable of analysis, a principle that has ultimately led McIntyre and Walker (2019, p. 61) to consider the notion of 'corpus-informed stylistics' – humans using statistical results from the machine processing of large text corpora to support human judgements on the validity of qualitative, intuitive accounts of shorter texts, accounts that 'might otherwise be fairly subjective claims'. A computer can act as a tool for creating and processing a corpus, but it cannot, as Smith hoped, do the criticism. Throughout, he acknowledges the central role of the human researcher in guiding a computer through the critical process:

Mark Boardman :: m.boardman@hud.ac.uk :: markboardman@outlook.com :: +44 7780 515377
Fellow in English Language and Linguistics at The University of Huddersfield

> 'Since the computer can deal only with formal relations among characters, words, or other segments, the researcher must provide all concepts of "meaning"; this is usually done through a system or systems of categories.' (p. 332)

Because twenty-first century computers are direct descendants of Turing machines (essentially highly complex binary switching devices) computational processing is not a good fit for criticism, or for anything approaching human thought; see Wells (2006, pp. 199-200) for a discussion of how neuroscientific research seems to bear out this observation.

The development of evidence based qualitative stylistics, stylistics as science, has led to sometimes acrimonious exchanges between academics as to the relative merits of pursuing objectivity in text analysis, based on the extent to which a human observation on textual dynamics can be precisely replicated by different human analysts using the same methodology, and on the extent to which the same observation could be proven wrong (falsified) – also by an analyst using the same methods – features of scientific methodology popularised by Boyle (1671) and Popper (1959, 2002) respectively. There has been a consequent push towards statistically based judgements in stylistics, and a reaction against linguistically driven judgements in literary study, the latter motivated partly by a wish to retain the perceived value of subjective judgements within literary criticism. Exchanges have been sometimes protracted and unhelpfully emotional. Central to the difference of opinion has been the issue of perceived objectivity – which is perceived in turn to be supported by statistical evidence. Mackay (1996) is a representative sample from an English Literature teacher's response to that stance:

> 'To show that a writer 'favours' a specific type of word, one would be required to identify all the situations where such a word is used in the text(s) in question and then show that other, different types of words could have been used but were not; otherwise it could be argued that these words were simply the words she used because she was referring to situations described by these words. (It might be argued that a farm worker is more likely to use words like 'fodder' or 'cow shed' than is a lawyer, but we would not be right to say that farm workers 'favour' these words.) A word count by itself would prove nothing because words are not definable in numerical terms. Unless the argument were over a specific number of instances of a word, type of word, or whatever, where accurate computation would settle the matter, there is no statistical evidence that could be adduced to clinch the argument because the argument centres not on numbers but on the interpretation of language – in this case, the meaning of the word 'favours'.' (p. 83)

Mackay is describing a by then well established tenet in qualitative stylistics which seeks to validate personal stylistic judgements using statistical data. This tenet has come to form the basis of the sub-discipline known as 'corpus stylistics': see McIntyre (2017); McIntyre and Walker (2019). Short, Freeman, van Peer and Simpson (1998, February) published a reply to Mackay's article in which they sought to contextualise the concept of objectivity within qualitative stylistics:

> '... stylisticians do not delude themselves into thinking that they provide irrefutable analyses. They know they can be wrong, and make a point of making clear the grounds for their views, and how they could be falsified, precisely so that others can challenge and test them, thus advancing our knowledge about texts and how we understand them. For a stylistician, then, being objective means to be detailed, systematic and explicit in analysis, to lay one's interpretative cards, as it were, clearly upon the table. If you believe that the number of interpretations that a text can hold is not indefinitely large (see Alderson and Short [1988] and Short and van Peer [1988] for empirical evidence to support such a view), then interpretative argumentation and testing will have to depend not upon something as unreliable as rhetorical persuasion, but on analysis of the linguistic structure of texts in relation to what we know about the psychological and social processes involved in textual understanding. This is what stylistics has traditionally involved. Of course, as we pointed out in section 2 of this paper, we cannot expunge our personal response from our analyses, and would never want to. Like the natural and social scientists, we are human analysts, not machines.' (p. 46)

Key statements here are that 'we are human analysts, not machines', along with an implied assertion that 'rhetorical persuasion' is 'unreliable' in that it cannot form the sole basis of valid text analysis methodology. The problem with this view, for a literary critic, is that the academic disciplines traditionally known as English and English Literature, as defined by Eagleton (1993, 1996, 2008, pp. 15-46) are historically founded on cultural rhetoric and have historically resisted attempts to incorporate within them core methodologies derived from any systematic examination of textual form. Given this context, Smith (1978) was bold to suggest a merger between computers and humans for performing literary critical tasks. Boardman (2004) observes in relation to computers co-operating with humans on language related tasks:

Mark Boardman :: m.boardman@hud.ac.uk :: markboardman@outlook.com :: +44 7780 515377
Fellow in English Language and Linguistics at The University of Huddersfield

> 'To adapt a quote from the first 'Terminator' movie, you cannot bargain with them, or plead with them; they do not feel pity or remorse or fear, and they absolutely will not stop until they are told to in the right way. If a computer appears to understand natural language, it is because it has been programmed with some very sophisticated instructions which tell it to respond in specific ways according to the data it has received. The same context, and the same data, will always produce the same response. Humans are less reliable.' (p. 90)

For a stylistician, this represents a capable helper that can perform textual processing tasks inordinately faster and with potentially fewer errors than a human, allowing the human to make the transition to a relatively conservative qualitative account of textual dynamics arising from replicable computational data, because machines are not analysts; but for a literary critic it represents built-in limitations on the types of commentaries that a critic is permitted to provide on textual data and at the same time have those commentaries regarded as academically 'reliable'. Siemens (2002), citing Machan (1991) in a re-evaluation of Smith's paper, regards posited activities labelled 'Higher Criticism' and 'Lower Criticism' as relevant in this respect. The latter is defined as factually verifiable detail surrounding a text, such as context of production, as well as computer-processible linguistic surface form. 'Higher Criticism' refers to interpretive commentary intended to offer rhetorically argued (but textually derived) opinions as to the cultural, aesthetic and psychological mechanisms of signification deemed to operate within a literary text. This is the critical domain traditionally inhabited by literary critics, but there has so far been no widely accepted method of transitioning to this domain from computationally derived data. The role of corpus processing in the work of stylisticians is becoming ever more routine and mainstream in helping to derive qualitative judgements on textual data, as illustrated by McIntyre (2017) in the concluding comments of his talk aiming to provide a definition of the emerging sub-discipline of corpus stylistics (Figure 1):

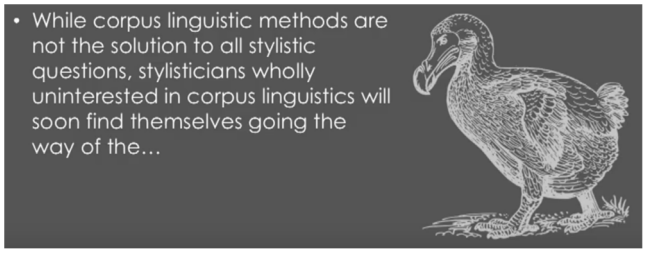

Figure 1: screen clipping from a PowerPoint presentation used by McIntyre (2017)

Currently deployed quantitative methodology within corpus stylistics can license Machan's qualitative 'Lower Criticism' in ways that allow stylisticians and perhaps some literary critics to remain comfortable, but before 'Higher Criticism' can be developed from computational data in ways acceptable to stylisticians and literary critics, we need a method that allows computers to license subjectivity in qualitative human analysis. This thought is encapsulated by Smith (1978) in the final paragraph of the article cited at the start of this section:

> 'Having translated the substantive hypothesis into functional terms and having used the computer to gather and display information and to explore various structural relations, it is then incumbent upon the critic to assimilate this information, to place it in context, and to synthesize his "interpretation." Obviously, the computer can only strengthen, not replace, his critical judgment. The final results of the inquiry should be expressed, once again, in the vernacular of the profession. To do this, the critic must translate in reverse the relations, patterns, and structures he has discovered on the functional level back into meaningful critical assertions. The computer should recede into the background leaving behind the unencumbered thesis, but a thesis that rests firmly on a body of specifiable assumptions and demonstratable textual relations. It is this joining of the deductive, critical response of the researcher with the empirical methodology of the computer that makes it possible to envision a science of literary criticism that is powerful but not reductive, sensitive but not simplistic.' (p. 354)

'Powerful', 'sensitive', 'reductive' and 'simplistic' are emotive terms – but they do express succinctly a so far the largely intractable issue of whether subjective human observations can form part of a critical process deemed to fulfil criteria associated with rigour, replicability and falsifiability.

Proposals for critical corpus construction and 'CCML'

'Corpus Critical Markup Language' (hereafter 'CCML') is defined by this paper as a corpus annotation system based on and conforming to XML (Extensible Markup Language) as specified by The World Wide Web Consortium (W3C) (2008). CCML is the basis of corpus construction and annotation within the critical method defined, specified and described by this paper and hereafter known as 'Corpus Criticism'. CCML requires that corpora be both 'well formed' and 'valid' XML documents as defined in the W3C specifications cited above. Additionally, Corpus Criticism refers to works selected for inclusion in literary corpora as 'documents' rather than texts; literary corpora created using CCML are also referred to as 'documents'. CCML corpora are wholly contained within the tag (or XML 'element') <CriticalCorpus> which has an XML namespace qualified by the URI: http://corpuscriticism.com/CCMLSchema/. The same URI also points to the server location where XML schema files used to validate CCML corpora are stored.

Concepts discussed by Berners-Lee and Hendler (2001) have redefined and reshaped the social mediation of written language in first two decades of the twenty-first century, including the social mediation of literary works. The processing of XML document metadata has come to define and underpin the daily lives of everyone on the planet – for example in key areas such as online research, commerce, banking, written communication, social media and control of national infrastructures – so it seems logical to make literary corpora conform to clearly defined rules of XML well-formedness and validity, thus making them cross-compatible, in terms of descriptive markup, with any current or future corpus processing software. It might be objected here that the TEI Consortium (2021, April 9) have already established such a system, but TEI is a highly complex system with a very steep learning curve and, unlike CCML, is designed to encompass multiple contexts in which any electronic text might be stored or interchanged, rather than being specifically optimised for productive cross-fertilisation between subjective literary interpretations and computational document processing – as is the case with CCML. Corpus Criticism treats critical writing as an act of document restructuring, and reflects this in its corpus construction and document encoding principles. Central to Corpus Criticism is the principle that the author and constructor of a critical literary corpus needs to have precise control over the corpus metadata, such that the corpus is parsed as well formed and valid by a recognised XML validation program or tool.

Below is a screenshot of the start of CCML corpus "CCML-Dickinson-Boardman-Fascicle-Poems" authored and constructed by me as part of the primary data for my PhD research investigating how persona and grammatical agency might intersect with prototypical linguistic and thematic evidence of a modernist consciousness (Brown, 1989; Sotirova, 2013) in Emily Dickinson's 818 fascicle poems. The poems are sequenced according to Miller's (2016) classification. Namespace and schema location are declared according to principles outlined above. The corpus was constructed and edited using Microsoft Visual Studio Community Edition (2021) which performs real time checks on the XML well-formedness and validity of the document, displaying the results of these checks at the bottom of the window (Figure 2).

Mark Boardman :: m.boardman@hud.ac.uk :: markboardman@outlook.com :: +44 7780 515377
Fellow in English Language and Linguistics at The University of Huddersfield

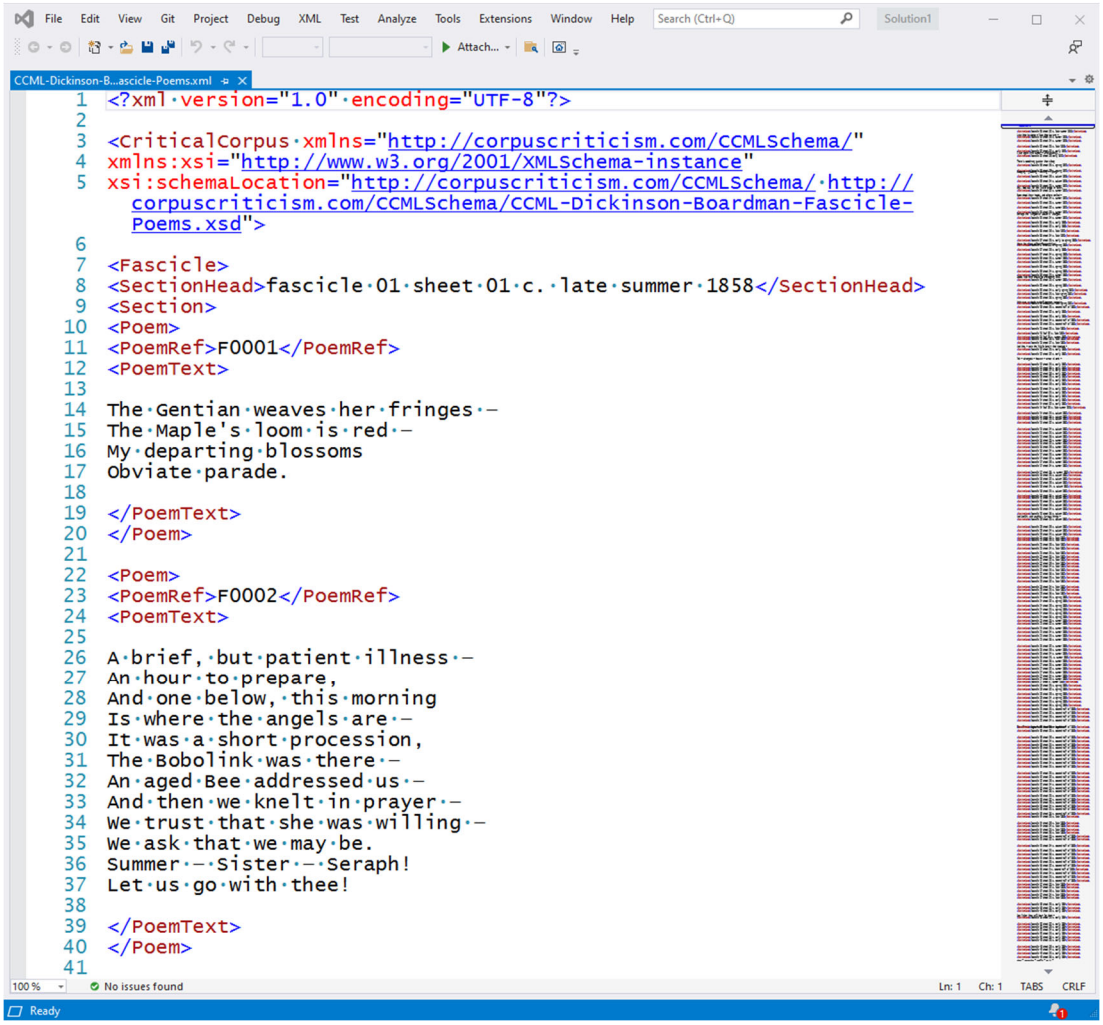

Figure 2: extract from the opening of 'critical corpus' "CCML-Dickinson-Boardman-Fascicle-Poems"

In order to test for stylistic markers, the next step is to process the corpus using one of many available corpus linguistic software packages. Most of these are primarily lexical and semantic in focus. As was outlined in Boardman (2019), my PhD research involves querying corpora for syntactic features as potential markers of agency, which led me choose NooJ (Silberztein, 2020) as a corpus processor. In order to link these markers in a reliably replicable way with literary tropes (themes) that might form the basis of Machan's (1991) 'Higher Criticism' – often deemed too subjective to be central to qualitative stylistics – I propose the following workflow elements in the construction and processing of CCML corpora.

A 'Seed Critique' is defined by this paper as a short subjective reading of a section of the corpus (in this case a reading of one poem from the 818 fascicle poems) – that is typically around 200 to 300 words in length and contains observations on linguistic markers and literary tropes that the critic intends to test computationally against the whole corpus. The seed critique is given a unique name, in this case 'FragmentedSelf', which is used to define an XML tag within the CCML namespace and schema structure. The full wording of the seed critique is reproduced and embedded within the comments of the relevant XML schema file, using the identification format reproduced in the next section of this paper. If a given 'critical corpus' is required to be validated against a schema stored at the location indicated by the URI http://corpuscriticism.com/CCMLSchema/, then that critical corpus will always need to be constructed according to strictly replicable criteria – as files at this URI cannot be modified freely, only downloaded and viewed: for schema files to be modified or added at the declared location for CCML, permission and agreement would need to be sought.

Seed critique: a proto-modernist deconstruction of Poem F0273

The seed critique modelled in the following workflow is a deconstructive modernist reading of poem F0273, from my corpus "CCML-Dickinson-Boardman-Fascicle-Poems" validated against a declared schema of the

same name located at the CCML URI named above. The research question that the seed critique attempts to answer in a brief, notional and subjective way is whether poem F0273 displays a prototypical Mrs Dalloway-like fragmented consciousness according to a combination of critical concepts put forward by Brown (1989), Sotirova (2013) and Derrida and Butler (1967, 2016), indicative of a disempowered persona experiencing social and psychological alienation characterised by the literary movement known as modernism. The entire corpus will then be queried for occurrences of this concept ('FragmentedSelf') and a computational process used to isolate intersections between a proto-modernist fragmented self and syntactic markers of disempowerment.

Any literary critical school or method, or combination of methods, could be used to inform the seed critique, but deconstruction feels like a natural fit for Dickinson and a potentially modernist consciousness, reflecting as it does the illusory nature of binary meaning, and arguing, as it does, for writing as a communicative mode with forms and mechanisms of signification distinct from those of speech. Although the first person persona in many of Dickinson's poems ostensibly mimics speech, analysing them in terms of spoken norms quickly leads to apparently insoluble pragmatic conundrums, so the classic Saussurean (1916, 2013) position on the primacy of speech in the creation of linguistic meaning, with writing represented as a secondary manifestation of speech, seems inadequate for the linguistic intricacies of modernist writing:

> 'For if one leaves out of account that multitude of movements required to actualise it in speech, each sound pattern, as we shall see, is only the sum of a limited number of elements or speech sounds, and these can in turn be represented by a corresponding number of symbols in writing. Our ability to identify elements of linguistic structure in this way is what makes it possible for dictionaries and grammars to give us a faithful representation of a language. A language is a repository of sound patterns, and writing is their tangible form.' (p. 67)

There is a core of truth in this for the founding principles of NLP (see Kurdi (2016) and (Kurdi, 2017) for a very full explanation of NLP principles), but Derrida's (1967, 2016) response to Saussure's position encapsulates how limiting Saussure's perspective is from a literary critical standpoint:

> 'That the "imprint" is irreducible also says that speech is originarily passive, but in a sense of passivity that all intramundane metaphors would only betray. This passivity is also the relationship to a past, to an always-already-there that no reactivation of the origin could fully master and awaken to presence. This impossibility of reanimating absolutely the manifest evidence of an originary presence takes us back therefore to an absolute past. That is what authorized us to call trace that which does not let itself be summed up in the simplicity of a present. It could in fact have been objected that, in the indecomposable synthesis of temporalization, protention is as indispensable as retention. And their two dimensions are not added up but the one implies the other in a strange fashion. What is anticipated in protention does not disjoint the present any less from its self-identity than does that which is retained in the trace. Certainly.' (p. 71)

Derrida's opaque written expression is a barrier, but the critically usable essence that can be extracted is that the self fragments due to an irreconcilably fluid relationship between the signifier and the signified, and that contrary to Saussure's position, writing has its own systems of signification distinct from those of speech. Al-muttalibi (2018) productively applies some concepts from deconstruction theory to selected Dickinson poems, largely based on binary lexical and pairs and sematic domains, but it is rather the *non-binary* illusive nature of the self that makes Brown's (1989) fragmented ('fragmentary') self an apt critical bedfellow for Derrida:

> '... I view the Modernist representation of selfhood as characteristically deconstructive. What is finally meant by 'self-fragmentation' can only emerge in terms of the developing argument. But the general implication is evident. Works such as *Ulysses*, *The Waste Land*, *The Waves*, the Pisan *Cantos* and *Four Quartets* represent, through experimental means, a selfhood which is pluralist, heterogeneous and discontinuous.' (pp. 1-2)

> 'The fragmentary self is rendered as selfhood where all parts and aspects are of the same interest; where everything is equally valid. And yet everything flows, too, in a forward-moving current of rhythmical energy where repetition also provides a sense of almost musical patterning. So the very discourse of Mrs Dalloway exemplifies its message concerning authentic selfhood. Selfhood is not unitary, but multiple, changeable, heterogeneous. And this truth is conveyed by a language which has released itself from the normal tyranny of 'rational' narrative progression and exploits the possibility of quasipoetic effects to express that 'incessant shower of innumerable atoms' which is the self's life.' (p. 107)

Mark Boardman :: m.boardman@hud.ac.uk :: markboardman@outlook.com :: +44 7780 515377
Fellow in English Language and Linguistics at The University of Huddersfield

This perspective is developed by Sotirova (2013) within the context of twenty-first century literary stylistics:

> 'The so-called revolution in the presentation of consciousness accomplished by Modernist writers, then, lies not primarily, and not exclusively, in the dismantling of the illusory coherence of thought and perception for the sake of a verisimilar transcription of the inchoateness of the stream of consciousness. It is – from the linguistic or stylistic viewpoint – the artistic desire to express simultaneously and concurrently the life-abstracted patterns of fragmentation and wholeness. If the fragmented grammatical and discourse structures can be read as the index of fragmented social structures, and fragmented selves, the redemptive power of the art of these experimental writers, is, from the philosophical point of view also, unfolded as a dialogic technique of literary style.' (pp. 195-196)

There is a convincing agenda here for attempting to connect features of linguistic surface form to discourse markers that might predictably instantiate the modernist trope that Brown labels 'fragmentary self'.

The surface form text of poem F0273 is reproduced below:

> 'You see I cannot see — your lifetime —
> I must guess —
> How many times it ache for me — today — Confess —
> How many times for my far sake
> The brave eyes film —
> But I guess guessing hurts —
> Mine — get so dim!
>
> Too vague — the face —
> My own — so patient — covets —
> Too far — the strength —
> My timidness enfolds —
> Haunting the Heart —
> Like her translated faces —
> Teazing the want —
> It — only — can suffice!'

A deconstructive reading of the poem, based a the posited binary conceptual pair of 'self-as-agent' opposed to 'self-as-disempowered' forms the basis of following CCML seed critique:

> 'The poem sets up the first person side of an implied dialogue between two personae manifested linguistically as first and second person pronominal and possessive adjectival paradigms. No name references or references to social relationships are used. Syntactic predictive mechanisms within the poem are driven by the following verbal lexemes: 'see', 'guess', 'ache', 'film', 'get', 'covet', 'enfold', 'haunt', 'translate', 'teaze', 'suffice' – most either directly or metaphorically abstract. Because conventional punctuation is not used, there is some perceived ambiguity as to the segmenting of verbs and predicates. For example, it is possible that both instances of 'see' in line 6269 are intransitive, or that the noun phrase 'your lifetime' in the same line is a direct object of the second 'see'. Alternative syntactic processing could read 'your lifetime' either as an isolated noun phrase or as a direct object of 'must guess' in line 6270 as part of an OSV sequence. It is this type of syntactic ambiguity that leads to an overall perception of uncertain personal agency – as though attempts to assert agency within the relationship are characterised by repeated unsuccessful attempts to second guess the other person's feelings. The nature of any perceived psychological processing behind the linguistic surface forms is illusory. Lexemes 'enfold', 'haunt', 'translate', 'teaze' represent a sequence of emotional parries that, it is implied, the second person persona may have found it difficult to process, concluding with 'suffice' as an undefined emotional compromise. These multi-valent surface form features point to an unstable relationship between signifier and signified in the Derridean (1967, 2016) sense, indexical of multiple selves engaged in internal psychological as well as external social fragmentation. Such signifier-signified instability is potential evidence of Brown's (1989) concept of 'language which has released itself from the normal tyranny of 'rational' narrative progression' as well as Sotirova's (2013) concept of 'fragmented grammatical and discourse structures' which 'can be read as the index of fragmented social structures, and fragmented selves'.' (Seed critique written by Mark Boardman, 2021.)

The seed critique is then embedded within the comments of the *.xsd schema file that is being used to validate the corpus:

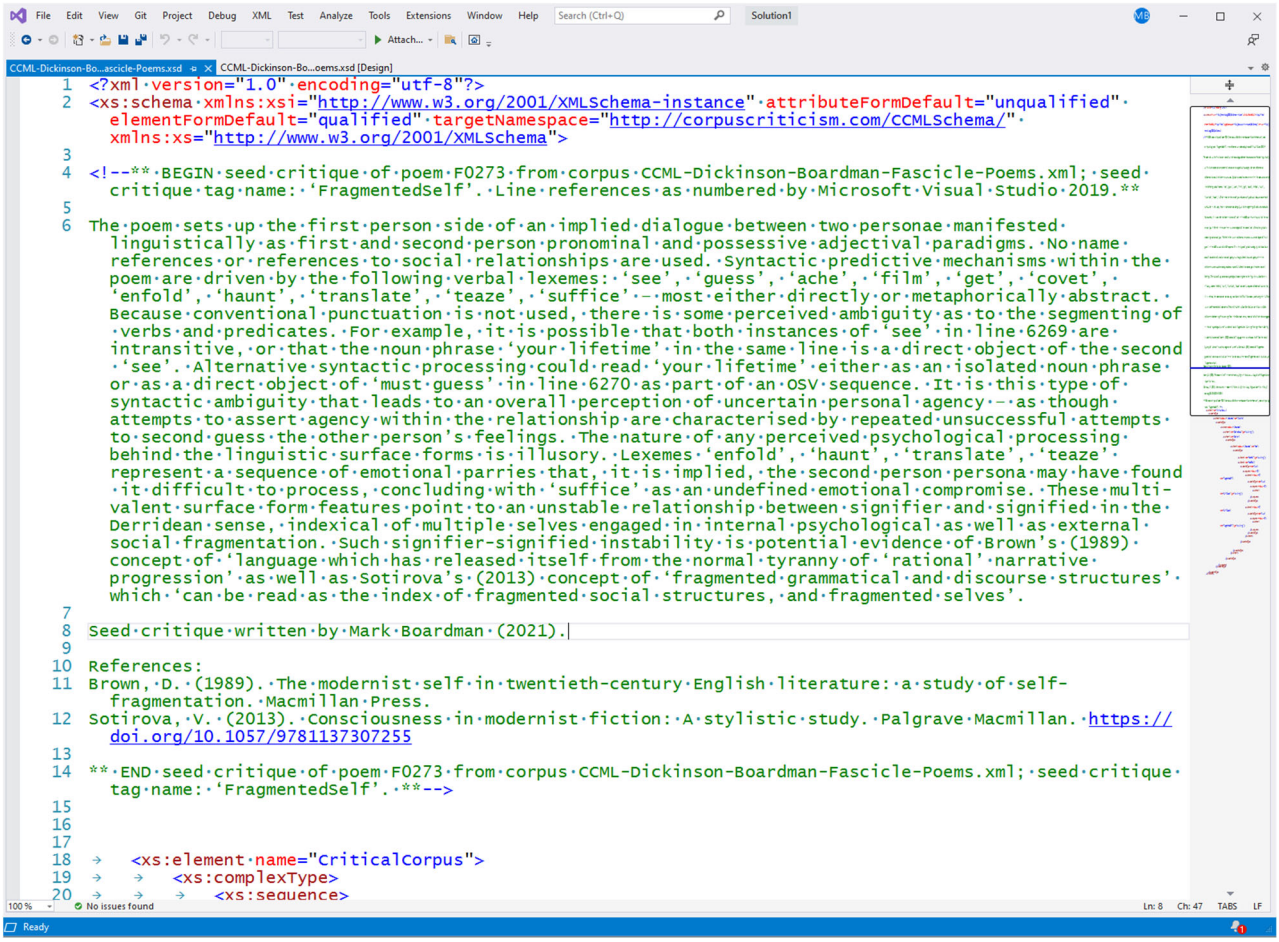

Figure 3: opening of CCML schema file "CCML-Dickinson-Boardman-Fascicle-Poems.xsd" used to validate CCML corpus "CCML-Dickinson-Boardman-Fascicle-Poems.xml"

Text enclosed in <!-- --> bracket pairs is not processed by XML parsing engines, but is placed in the schema file as a reference for the author of the Critical Corpus as to the precise definition of the XML tag <FragmentedSelf>. The next step is to enclose poem F0273 within the <FragmentedSelf> tag (Figure 4). Finally, the author makes a critical judgement, based on the features identified in the seed critique, as to where else in the corpus there is linguistic and literary critical evidence of <FragmentedSelf>, and encloses all such identified sections of the corpus's surface form stream (which must, according to the constraints of the schema, be contained inside the <PoemText> element) within the tag <FragmentedSelf>. Figure 5 shows how the <FragmentedSelf> element is defined by the schema as part of a nested hierarchy.

This act of document restructuring is central to Corpus Criticism theory. Emily Dickinson's 818 fascicle poems are no longer typeset surface form printed on paper or displayed for an e-reading platform: the surface form now has an intrinsic structural and critical relationship with the corpus metadata. Another critic would obviously not do the restructuring in the same way, but as long as any critic follows the procedures outlined above in constructing a well-formed and valid XML corpus, the methodology of this phase of the critical process will always be replicable. The next computational phase of a Corpus Critical reading is to set up and implement parameters for a chosen corpus processing tool to query the corpus, the results of which will be used to inform a more detailed qualitative analysis of the primary data. These procedures will be explained fully in "From token to exegesis part 2", but they are summarised in the final section of this paper.

Mark Boardman :: m.boardman@hud.ac.uk :: markboardman@outlook.com :: +44 7780 515377
Fellow in English Language and Linguistics at The University of Huddersfield

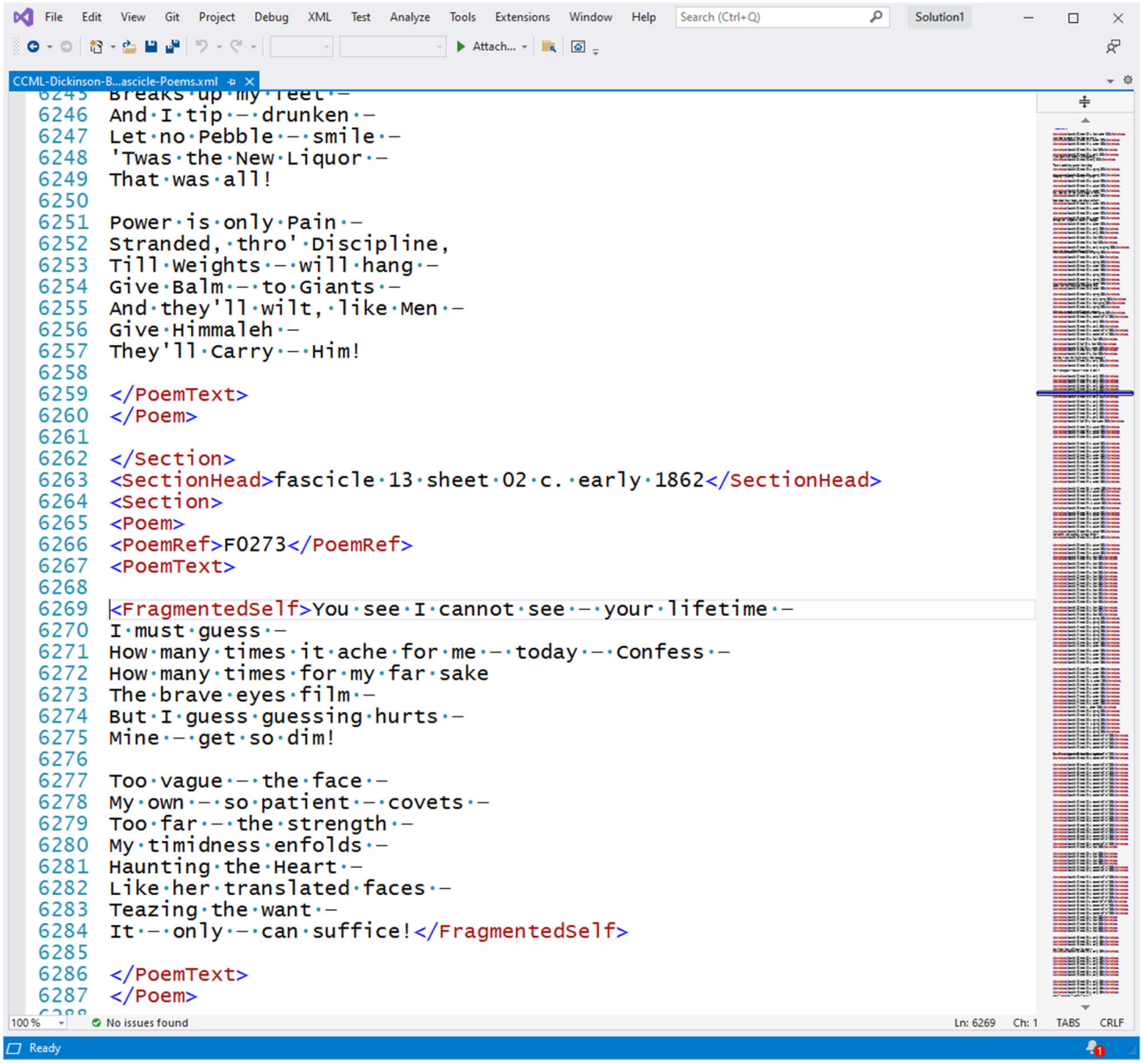

Figure 4: poem F0273 is defined with the XML element <FragmentedSelf> which the schema will only allow to be enclosed with in the element <PoemText>

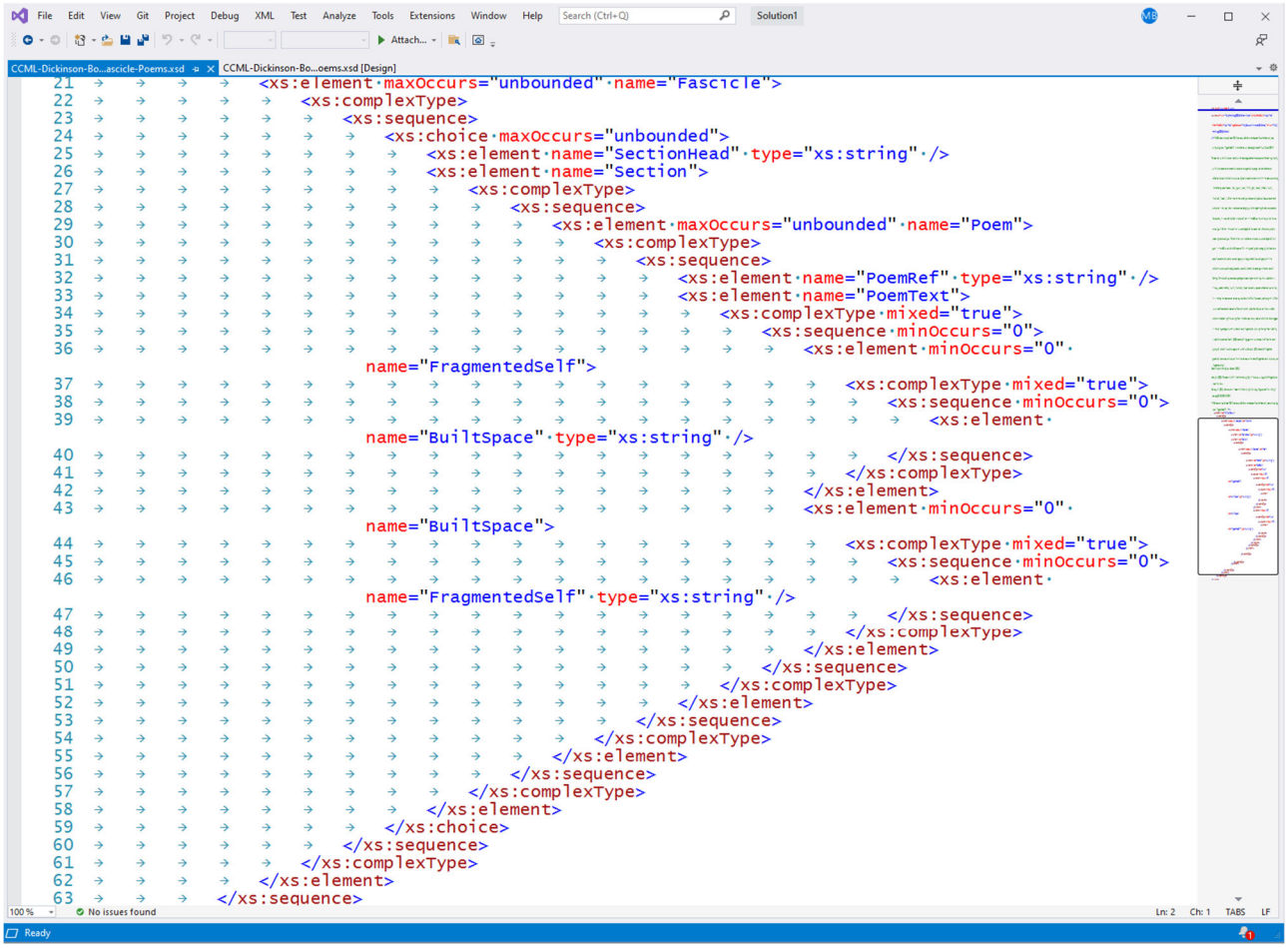

Figure 5: how the <FragmentedSelf> element fits into the schema within a nested hierarchy

In part 2

NooJ (Silberztein, 2020) is a rule based NLP development package and corpus processor which takes its linguistic processing parameters from user-developed files containing morphological, syntactic, lexical and semantic instructions, rather than deriving probabilistic patterns from reference corpora. NooJ can import XML structure and render XML elements as dictionaries. Thus, NooJ can be instructed to treat CCML seed critique elements in a corpus as structurally and semantically meaningful. If this procedure is combined with methods outlined in Boardman (2019) for writing NooJ grammars to look for *syntactic* features of agency that interact with perceptions of persona, it is possible to write a series of NooJ grammars that will provide concordance outputs indicative of whether the critical concept <FragmentedSelf> shows any typicality in the linguistic surface form. In turn, it should be possible to develop a view, through argumentation, of the linguistic, aesthetic and cultural behaviour of <FragmentedSelf> as a discourse feature. Part 2 will aim to consolidate the sense in which Jeffries and Walker (2020) have already used the phrase 'corpus critical' – the principle of linking of concordance outputs to qualitative critical theory. However, Corpus Criticism, as introduced in this paper, proceeds in a different order, beginning with a notionally derived subjective reading and seeking to expand and stabilise this reading into a more powerful qualitative interpretation via a middle computational stage. Corpus Criticism is a three-way hybrid that gives Literary Criticism, Corpus Stylistics and Natural Language Processing equal status.

Mark Boardman
July 2021

References

Al-muttalibi, A. Y. (2018). A Deconstructive Reading of Dickinson's Poetic Texts. *Advances in Language and Literary Studies*, *9*, 144. https://doi.org/10.7575/aiac.alls.v.9n.6p.144

Berners-Lee, T., & Hendler, J. (2001). Publishing on the semantic web. *Nature*, *410*(6832), 1023–1024. Nature Publishing Group. https://doi.org/10.1038/35074206

Boardman, M. (2004). *The Language of Websites*. Routledge.

Boardman, M. (2019). Grammatical agency and ironic persona in Emily Dickinson: an interdisciplinary corpus originated study. *Online Proceedings of the Annual Conference of the Poetics and Linguistics Association*. PALA. https://doi.org/10.6084/m9.figshare.14178605

Boyle, R. (1671). *Some Considerations Touching the Usefulnesse Of Expirmental Naturall Philosophy* (Vol. 2). Oxford.

Brown, D. (1989). *The modernist self in twentieth-century English literature: a study of self-fragmentation*. Macmillan Press.

Chomsky, N. (1965, 2015). *Aspects of the Theory of Syntax* (50th Anniversary Ed.). MIT Press.

de Saussure, F., & Harris, R. (1916, 2013). *Course in General Linguistics* (R. Harris, Trans.). Bloomsbury. https://amzn.to/2U0b2Ld

Derrida, J., & Butler, J. (1967, 2016). *Of Grammatology* (G. C. Spivak, Trans.). (Fortieth Anniversary Ed.). John Hopkins University Press. https://amzn.to/2SY7QzJ

Eagleton, T. (1993, 1996, 2008). *Literary theory: an introduction* (Anniversary Ed.). Blackwell Publishing.

Jeffries, L., & Walker, B. (2020). Austerity in the Commons: A corpus critical analysis of austerity and its surrounding grammatical context in Hansard (1803– 2015) In Power, K., Ali, T., & Lebdušková, E. (Eds.), *Discourse Analysis and Austerity: Critical Studies from Economics and Linguistics*. Routledge.

Kurdi, M. Z. (2016). *Natural language processing and computational linguistics: 1, Speech, morphology and syntax*. iSTE. https://doi.org/10.1002/9781119145554

Kurdi, M. Z. (2017). *Natural language processing and computational linguistics: 2, Semantics, discourse and applications*. ISTE. https://doi.org/10.1002/9781119419686

Machan, T. W. (1991). Late Middle English Texts and the Higher and Lower Criticisms In Machan, T. W. (Ed.) *Medieval Literature: Texts and Interpretation*. Center for Medieval and Early Renaissance Studies (State University of New York at Binghamton).

Mackay, R. (1996). Mything the point: a critique of objective stylistics. *Language & Communication*, *16*(1), 81-93. Pergamon.

McIntyre, D. (2017, July 28). Just what is corpus stylistics? [Video]. University of Birmingham. YouTube. https://www.youtube.com/watch?v=X7vuRzvQ0nQ

McIntyre, D., & Walker, B. (2019). *Corpus stylistics: theory and practice*. Edinburgh University Press.

Microsoft Corporation (2021). *Microsoft Visual Studio Community 2019* (Version 16.9.6) [Computer software]. https://visualstudio.microsoft.com/downloads/

Miller, C. (Ed.). (2016) *Emily Dickinson's Poems As She Preserved Them*. The Belknap Press of Harvard University Press.

Popper, K. (1959, 2002). *The Logic of Scientific Discovery*. Routledge.

Short, M., Freeman, D. C., van Peer, W., & Simpson, P. (1998, February). Stylistics, criticism and mythrepresentation again: squaring the circle with Ray Mackay's subjective solution for all problems. *Language and Literature: International Journal of Stylistics*, *7*(1), 39-50. https://doi.org/10.1177/096394709800700103

Siemens, R. G. (2002). A New Computer-Assisted Literary Criticism? *Computers and the humanities*, *36*(3), 259-267. Kluwer Academic Publishers. https://doi.org/10.1023/A:1016134426453

Silberztein, M. (2016). *Formalizing natural languages: the NooJ approach*. ISTE. https://doi.org/10.1002/9781119264125

Silberztein, M. (2020). *NooJ* (Version 7.0 b20200621) [Computer software]. http://www.nooj-association.org/downloads.html

Smith, J. B. (1978). Computer Criticism. *Style (University Park, PA)*, *12*(4), 326-356. University of Arkansas.

Sotirova, V. (2013). *Consciousness in modernist fiction: A stylistic study*. Palgrave Macmillan. https://doi.org/10.1057/9781137307255

TEI Consortium. (2021, April 9). *The TEI Guidelines* https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

The World Wide Web Consortium (W3C). (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)* https://www.w3.org/TR/2008/REC-xml-20081126

Turing, A., Ford, K., Gylmour, C., Hayes, P., Harnad, S., & Saygin, A. P. (1950, 2008). Computing Machinery and Intelligence In Epstein, R., Roberts, G., & Beber, G. (Eds.), *Parsing the Turing Test* (pp. 23-70). Springer. https://doi.org/10.1007/978-1-4020-6710-5

Wells, A. (2006). *Rethinking cognitive computation: Turing and the science of the mind*. Palgrave Macmillan.