Affordances and limitations of algorithmic criticism

Verhaar, P.A.F.

Chapter 3

Current state of literary informatics

3.1. Introduction

Ever since computers were given the ability to process alphanumerical characters, scholars have experimented with the numerous ways in which the digital medium can query and manipulate works of literature. To a large extent, the appeal of computation lies in the unequalled speed with which information can be extracted from texts. Once the rules for the identification of features of interest have been implemented in an application, the actual time needed for the execution of this algorithm is, in most cases, negligible. Consequently, it becomes practicable to apply a set of algorithms iteratively to text bases of thousands or of millions of documents.[202] In addition, while the particular ways in which texts are processed by human scholars is often influenced, to a higher or lesser degree, by the scholar's mood or by levels of concentration, computers are incapable of fatigue, and apply the rules that are specified in an algorithm with unrelenting rigour and consistency to each text in the corpus.

This chapter discusses current scholarly practices in the field of literary informatics. As was discussed, many of its methods are based on text mining technologies. While text mining and NLP are rich and burgeoning areas of research, the aim of this chapter is not to describe the achievements and technical challenges in these areas as such. The scope remains limited to the tools and the concepts that have been developed or adopted by humanities scholars for the investigation or the clarification of literary texts. The main aim is to describe the concepts and principles that underlie text analysis tools, and to arrive at a better understanding of the methodology and the epistemology of computer-based literary research.

Studies in the field of literary informatics frequently make use of standardised text analysis tools. Many examples of such applications can be found via online directories such as Tapor[203] and Dirt.[204] The nature of literary informatics research can partly be examined by considering the basic functionalities which are offered by these tools. Evidently, the tools that have been developed in research projects have not all been made available publicly as independent production level services. This is not always possible, because tools have sometimes been developed for very

[202] Gregory Crane, "What Do You Do with a Million Books?", in: *D-Lib Magazine*, 12:3 (2006).
[203] <http://www.tapor.ca/> (19 October 2013)
[204] <http://dirtdirectory.org/> (19 October 2013)

specific research goals, or for highly specialised data sets. The following section provides an overview of the general functionalities which are offered by these tools. The third section of this chapter focuses on the various ways in which data and tools have been used concretely within research projects, and on the various research questions that have been addressed using these tools.

3.2. Tools

3.2.1. Vocabulary

Many of the functionalities that can be offered by text analysis tools are based on counts of the words that appear within a text. Users can often upload texts in the plain text format, and the program can subsequently divide the text into smaller linguistic units, such as, for instance, its words or sentences. This preparatory process is generally referred to as "segmentation" or "tokenisation". Segmentation generally takes place on the basis of the spaces, punctuation marks and line breaks that occur in the text. Such notational conventions are currently used in virtually all of our written natural language texts, and they have been in use at least since the Carolingian Renaissance in the late 9th century. Scribes in the early Middle Ages introduced a number of rules aimed at rendering the ocean of words that was found in ancient *scriptura continua* in a more legible form. Innovations included the use of spaces in between words, the distinction between upper and lower case, and the insertion of punctuation to mark the end of a sentence.[205] On the basis of the notational conventions, which Feldman refers to as 'soft markup',[206] text mining applications can be developed for the recognition of units such as words, sentences or paragraphs. [207] The total number of words that are found are referred to as "tokens", and the unique words are called "types". Frequency lists, which count occurrences of types, form the basis for further statistical analyses.

If word segmentation takes place exclusively on the basis of the usage of spaces, this is arguably a rather crude method. Views may vary, for example, on what exactly constitutes a word. Brinton explains that there are several ways in which the boundaries of words may be determined. [208] He discusses a distinction between orthographic and semantic criteria. In the orthographic approach, a word is simply

[205] Paul Saenger, *Space between Words : The Origins of Silent Reading* (Stanford Calif.: Stanford University Press 1997), p. 10.

[206] Ronen Feldman, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, p. 3.

[207] Very few tools offer support for sentence segmentation. One tool which may be used is the Lingua::Eng::Sentence in the Perl programming language <http://search.cpan.org/~achimru/Lingua-Sentence-1.05/lib/Lingua/Sentence.pm> (10 October 2013)

[208] Laurel Brinton, *The Structure of Modern English a Linguistic Introduction* (Philadelphia: John Benjamins 2000), pp. 73–74.

defined as a string of characters delineated by spaces. Complications arise, however, in the case of hyphenated words, compound words or phrasal verbs. Compound words obviously form one semantic unit, but they are generally treated as separate types by tokenisation algorithms. John Kirk discusses an additional number of complexities. Scholars may have dissimilar views on whether words in different spellings, e.g. the difference between 'color' and 'colour' in British and American spelling, ought to be treated as different types. Dialect words "represent various pronunciation variants of the same lexical type".[209] Counts, furthermore, may be based either on the lemma, being the dictionary entry of the word, or on the variously inflected forms of words. One additional difficulty is that tools often ignore the distinction between characters in upper case and in lower case by default. In projects that study the religious nature of texts, for instance, the occurrence of capitals in masculine personal pronouns may be highly relevant. John Burrows also notes that words that are treated by the tokeniser as a single type may nevertheless have widely distinct meanings. Current methods seem unable to deal adequately with "truly polysemous words like "blue" where numerous literal meanings shade off into all sorts of metaphorical senses".[210] Since research projects are not always fully transparent with respect to the way in which they actually produce counts, Kirk stresses that the use of frequency lists invariably demands interpretation.[211] Frequency lists should routinely be treated with caution.

Word segmentation is complicated by the fact that the use of punctuation marks such as apostrophes and hyphens is generally unpredictable. The exact purpose of these characters often varies with the context and with personal stylistic preferences of authors. The identification of words and sentences in English texts is normally relatively easy for human readers who have a proficiency in that language, but, for computers, "dealing with hyphenated words, apostrophes, conventions of using single and double quotes, and so forth all require the programmer's attention".[212] Different text analysis tools may have implemented different rules for tokenising texts on the basis of spaces and punctuation marks, and such discrepancies evidently lead to different counts. Taporware, for instance, removes punctuation marks such as question marks, quotation marks and exclamation marks, but does not remove trailing and leading hyphens and asterisks. Voyant removes all punctuation marks, except for trailing quotation marks and hyphens. The Lexomics tool offers users the possibility to remove all punctuation, with the option, nevertheless, to retain hyphens or word-internal

[209] John M. Kirk, "Word Frequency: Use or Misuse?", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009, pp. 19–20.

[210] John Burrows, "Never Say Always Again: Reflections on the Numbers Game", in: Willard McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Cambridge: Open Book Publishers 2010, p. 26.

[211] John M. Kirk, "Word Frequency: Use or Misuse?", p. 33.

[212] Roger Bilisoly, *Practical Text Mining with Perl* (Hoboken, N.J.: Wiley 2008), p. 8.

apostrophes. While such differences may appear trivial, they have a direct impact on the counts of tokens, and, consequently, on the results of text analyses.

The resultant counts may be analysed in a variety of ways. A number of tools support the creation of a keyword in context list or a concordance. This entails, more specifically, a functionality in which occurrences of a given search term can be shown in combination with words that occur before and after this term. Users can usually specify the length of the fragments in such lists. Frequencies can also be shown as a distribution graph. Distribution refers to a process in which a full text is divided into segments, and in which information is provided about the frequency of a particular word within each of these segments. Tools may also offer a collocation display, in which a frequency list of the search term that was provided is shown in combination with frequency lists of the words that occur most frequently within a certain distance of the provided keyword. In co-occurrence searchers, text fragments can be retrieved in which two given words occur within a certain distance of each other. Users can generally indicate the number of other words that may occur in between these words. In all of these operations, the focus is exclusively on the frequencies of words. These operations fully disregard the original word order. Approach which ignore the original context of words are commonly described collectively as the "bag of words" model.[213]

3.2.2. Linguistic features

As literary writing typically employs a register of language that differs in a number of important ways from common colloquial language, systematic analyses of literary texts commonly focus on the distinctive ways in which narratives and ideas have been expressed. Analyses of the language can be based on manipulations of the unprocessed plain text, but the scope of textual analyses can be expanded considerably by employing software that can enrich the plain texts with detailed annotations about their linguistic characteristics. Data about the syntactic categories of words, for instance, can be produced by making use of part of speech (POS) taggers. Many of the taggers that are available today implement a method that was proposed originally by Eric Brill. His method entails a process in which syntactic categories are initially supplied by making use of a large lexicon which consists of words from the English language, together with their potential grammatical categories. Using this lexicon, the text can be tagged provisionally. The results of the first round of tagging are subsequently improved through a set of transformation rules, which can detect cases of incorrect categorisation. It can be stipulated, for instance, that certain combinations, such as a singular pronoun followed by plural verb, can never occur. Using such approaches, POS tagging can

[213] Roger Bilisoly, *Practical Text Mining with Perl*, p. 123. On the basis of the definition that is given, it must be decided that a KWIC list is not based on the bag of words model.

take place semi-automatically. If the default lexicon is not accurate for a specific corpus, the tagger needs to be trained. This means that users will need to tag a set of text manually, so that the application can generate a new lexicon and new transformation rules.[214] POS taggers have been developed by the Stanford NLP team[215] and within the OpenNLP project.[216] In the PERL programming language, the module Lingua::Eng:Tagger[217] can be used, and the NLP toolkit for Python similarly includes a POS tagger.[218] While POS taggers typically make use of algorithms and transformation rules, data on grammatical categories can also be provided solely by making use of a lexicon. One example is Docuscope.[219] It is a corpus, developed by Michael Witmore at the University of Wisconsin, of several million English words and phrases of words which have been associated manually with specific grammatical, semantic and rhetorical categories. The functional linguistic categories which are used in DocuScope are referred to as 'Language Action Types' (LATs). Docuscope matches the string in the central dictionary to one of the pre-defined LATs. The strings "I" and "me", for instance, are be labelled with the LAT 'FirstPerson'.

Many existing pre-trained POS taggers assume a regularised spelling and a predictable structure. They do not function accurately in all cases. Kaplan notes that the results of POS taggers are frequently unreliable when they are used to parse the often complicated syntax of poetry. Poems often have a syntax which is deliberately ambiguous, leading to a situation in which taggers can potentially assign multiple categories. In the case historical texts or texts in dialects, challenges are posed by the fact that spelling and syntax can vary along with different eras or different regions. Many of these challenges have been addressed in the *Metadata Offers More Knowledge* (MONK) project, which ran from 2007 to 2009 under the direction of John Unsworth. The aim of the project was to enable humanities scholars to engage in text mining on the basis of tools which were already in use within the field of corpus linguistics. As part of the MONK project, a vast data store has been created with approximately 2500 texts, which collectively contain more than 150 million words. The corpus covers texts from different geographic areas and different historical periods. The objective, nonetheless, was to let scholars query these texts in a uniform manner. To allow for such searches across dialects and across historical eras, all texts have been encoded using a specific variety of TEI-P5, which was referred to as TEI-Analytics. In this encoding,

[214] Eric Brill, "A simple rule-based part of speech tagger", in *Proceedings of the third conference on Applied natural language processing*, (Morristown, NJ, USA: Association for Computational Linguistics, 1992), p. 152.
[215] <http://nlp.stanford.edu/software/tagger.shtml> (28 May 2014)
[216] <https://opennlp.apache.org/> (28 May 2014)
[217] <http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm> (28 May 2014)
[218] <http://www.nltk.org/> (28 May 2014)
[219] <http://www.cmu.edu/hss/english/research/docuscope.html> (26 September 2013)

each individual word was connected to its lemma, and to its syntactic category. As part of the project, a new POS tagger, named Morphadorner, was developed.[220] Within the project, it was decided that "the tokenizer should not sunder what the typesetter has joined", and, following this logic, contracted forms such as "th'earth" or "nilt" were not expanded.[221] The expanded forms were provided, nevertheless, in the lemmatised version, which connects the word form that is found in a text to its dictionary form. The MONK datastore and the software that has been developed form important resources for literary scholars who seek to explore patterns and regularities in historical text collections, based on extensive linguistic data.

Next to grammatical categories, scholars may also be interested in using data on phonetic aspects. Words, when read aloud, obviously contain sounds, and literary devices are often based on a skilful use of such sounds. Data on phonetics are especially relevant for the analysis of poetry. Since the English language does not have a close correspondence between orthography and pronunciation, it is not possible to extract data about sounds directly.[222] Phonetic transcriptions may be produced, nevertheless, by making use of pronunciation dictionaries. This approach was followed in the development of the PoetryAnalyzer tool, which was created by David Kaplan at Princeton University.[223] Among other functionalities, PoetryAnalyser enables scholars to identify literary devices such as perfect rhyme and alliteration in a text. The detection of these features are based on prior phonetic transcriptions, which are made using the openly available Carnegie Mellon Pronunciation dictionary.[224] Other dictionaries are available as well, but, as Kaplan was mainly interested in poets from the United States, this specific resource was chosen because it offers data on the pronunciation of American English. The tool can transcribe the tokens found in the texts via a lookup in the dictionary. Kaplan conceded that such direct lookups produce a small margin of errors, since certain words in the English language, such as "record" or "minute", may have different pronunciations, depending on their syntactic function. No measures were taken to correct these errors, however. Texts may also contain proper nouns such as personal names and geographical terms, and these will in most cases not be listed in the dictionary. In addition, the method cannot deal adequately with diachronic and synchronic variations in pronunciation. The PoetryAnalyzer software is less effective for investigating texts by British poets, for instance.

[220] <http://morphadorner.northwestern.edu/>
[221] John Unsworth & Martin Mueller, *The MONK Project Final Report*, (2009), p. 6.
[222] Susan Hockey, *Electronic Texts in the Humanities: Principles and Practice*, p. 78.
[223] David Kaplan & D.M. Blei, "A Computational Approach to Style in American Poetry", in *Seventh IEEE International Conference on Data Mining*, (2007), pp. 553–558.
[224] <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (15 March 2013)

Phonetic transcriptions can alternatively be produced using text-to-speech software, of which MaryTTS is doubtlessly the most significant example.[225]

3.2.3. Semantic contents

The quantitative and statistical approach towards studying literature has frequently been under attack as a result of its perceived superficiality. Indeed, if the concern is predominantly with the symbols that carry meaning, rather than with the meaning itself, this is admittedly a rather shallow form of engagement. In this mode of textual analysis, the relationship to literary works appears to be equivalent to the manner in which texts are treated in disciplines such as bibliography, codicology, book history or library studies. While these latter fields take books or other carriers of information as their primary research objects, they do not necessarily concern themselves with the contents of these texts. These disciplines focus on data about texts, and may study the reception, the production processes, or the physical appearance of items. Studies in the field of algorithmic criticism, by contrast, aim to contribute to an improved understanding of the contents and of the more profound thematic concerns of literary works, and various attempts have been made to bridge the gap between the lexical codes and their semantic contents.

One of the ways in which the semantic aspects of texts may be uncovered is by making use of lexicons which map the text's tokens to pre-defined semantic categories. Examples of applications in which this principle is implemented include the Harvard General Inquirer,[226] the Linguistic Inquiry and Word Count (LIWC) tool,[227] the UCREL Semantic Analysis System (USAS)[228] and DocuScope.[229] The Harvard General Inquirer, firstly, consists of 182 categories, each of which are connected to an extensive list of words. The category "negative", for instance, contains over 2290 entries. The DocuScope tool, which was also mentioned in a previous section, can provide data both about grammatical features and about semantic aspects. Phrases such as "whilst," "when he," "as he", for example, are labelled with the LAT "Narrative Time". LIWC, thirdly, consists of general pre-defined semantic categories for the words that are used in a text. It uses categories for positive or negative emotions, mental processes, self-references, and causal words. The tool can therefore be used effectively for sentiment analysis. Fourthly, the UCREL Semantic Analysis System (USAS) application, which was developed at

[225] MaryTTS is as "an open-source, multilingual Text-to-Speech Synthesis platform written in Java", "originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University". It offers support for "German, British and American English, French, Italian, Swedish, Russian, Turkish, and Telugu". <http://mary.dfki.de/> (12 June 2013)

[226] <http://www.wjh.harvard.edu/~inquirer/> (12 June 2013)

[227] <http://www.liwc.net/> (12 June 2013)

[228] < http://ucrel.lancs.ac.uk/usas/> (12 June 2013)

[229] <http://www.cmu.edu/hss/english/research/docuscope.html> (12 June 2013)

the University of Lancaster, consists of 21 major domains, which expand into 232 more specific semantic field tags.

Next to such lexicon-based approaches, investigations of the semantic contents of texts can also be based on statistical processing of the vocabulary. Topic modelling is the prime example of this approach. It is a generic term which refers to a range of algorithms that can be used to determine the topics that occur in a text on the basis of the vocabulary used in individual documents. Topic modelling is performed most frequently on the basis of an algorithm which is known as Latent Dirichlet Allocation (LDA), which was first discussed in an article by David Blei et al.[230] LDA is also implemented in MALLET, a Java-based tool created at the University of Massachusetts-Amherst.[231] The tool takes a text collection as input, and produces a number of topics as a result. The topics that are returned by MALLET concretely consist of unnamed lists of words. The number of topics to be returned have to be supplied as a parameter before running the algorithm. Users of MALLET need to inspect and interpret the lists and provide topic labels themselves. The central idea of Topic Modelling is that documents contain topics, and that these topics manifest themselves through specific words. If certain words co-occur frequently in the same documents, in ways that have been defined as statistically significant, these are assumed to be about the same topic. MALLET does not only return the topics, but also compiles a list of the documents containing these various topics. In this way, Topic Modelling can also be used to cluster the documents that focus on specific topics.

3.2.4. Data analysis

Studies in the field of literary informatics concentrate to a large extent on a description of the style of literary texts. The term "style" is used ubiquitously within literary and linguistics research, and, as a result, it is open to many different interpretations and definitions. Based on an elaborate survey of the various conceptualisations of the term within German, French and Dutch traditions in linguistic and literary scholarship, Herrmann, van Dalen-Oskam and Schöch define the term "style" broadly as "a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively".[232] This conceptualization of "style" has also been adopted in this thesis. The formal features which are mentioned in the definition may refer to "linguistic features at the level of characters, lexicon, syntax, semantics", and also to "features going

[230] David M. Blei, Andrew Y. Ng & Michael I. Jordan, "Latent Dirichlet Allocation", in: *The Journal of Machine Learning Research*, 3 (1 March 2003).

[231] <http://mallet.cs.umass.edu/> (2 March 2014)

[232] J. Berenike Herrmann, Karina van Dalen-Oskam & Christof Schöch, "Revisiting Style, a Key Concept in Literary Studies", in: *Journal of Literary Theory*, 9:1 (2015), p. 44.

beyond the sentence, such as narrative perspective or textual macro-structure".[233] The authors argue that a text's style can be characterised via a careful examination of all occurrences of these formal features, and of the complicated ways in which these features can be combined.

On an abstract level, computational literary analyses begin with a quantification of some of the formal features which constitute the style of a text, making use of the applications which have been discussed in the previous sections. The process of quantification results in a series of variables and of associated values for these variables. Without further processing, it is usually difficult to see patterns within the frequencies of such style markers. To understand the nature of data sets more fully, it is often necessary to apply additional techniques. The statistical procedures which can help to model and to clarify existing data collections are commonly referred to as "statistical learning" techniques or "machine learning" techniques. James et al. make a distinction between supervised techniques and unsupervised techniques.[234] Examples of supervised techniques include Naive Bayes classification, logistic regression and Support Vector Machines.[235] Supervised machine learning techniques can be used, among other purposes, to classify texts. In the case of classification, researchers firstly need to assign labels or categories to the texts in a training set. Using these labels, the classification algorithms can construct a statistical model which may be used to make predictions about the categories of new and unlabelled texts. James et al. make an interesting distinction between "prediction" and "inference". In the former approach, statistical learning techniques are applied primarily to classify unlabelled texts. In the case of inference, the focus lies mainly on an examination of the formal properties which, according to the model that was created, are typical for the texts in specific categories.

In the case of unsupervised learning techniques, scholars do not supply prior information about the potential categories of texts. This second class of techniques aims to discover patterns, clusters or relationships within unlabelled data. Patterns can, in some cases, be found simply by sorting or filtering the data on the basis of a particular data value. A widely used and more advanced unsupervised machine learning technique is Principal Component Analysis (PCA). It is a form of multivariate analysis, which reduces the complexity of a multidimensional data set through the creation of a number of new composite variables which account for most of the variability of the original variables. These new variables are referred to as the principal components. By plotting a limited number of principal components, certain patterns can be explored in the global distribution of the data values.

---

[233] J. Berenike Herrmann, Karina van Dalen-Oskam & Christof Schöch, "Revisiting Style, a Key Concept in Literary Studies", p. 44.

[234] G. James et al., *An Introduction to Statistical Learning, with Applications in R* (Springer 2013), p. 1.

[235] Joachim Diederich (ed.), *Rule Extraction from Support Vector Machines* (Berlin: Springer 2008).

Diagrams in which PCA are visualised can disclose the words that occur in similar frequencies, or can indicate the texts which use very similar words. Differences between texts can additionally be characterised via the calculation of the Euclidean distance or of the cosine similarity. The distances between texts may be clarified in the form of a dendrogram. In such diagrams, the texts which are most similar form a single branch, and texts which display fewer similarities do not form a union until a much later stage. As such, the method provides a highly intuitive method for clarifying the differences and the similarities between texts.

In general, quantitative analyses of data about the stylistic properties of text can only clarify patterns of differences and similarities. In literary informatics research, such statistical comparisons are eventually used to answer questions which are more directly germane to literary criticism. As will be discussed in the next sections, the results of such procedures can be used, for instance, to compare texts written by different authors, speeches uttered by different literary characters, texts written by male and by female authors or texts from different literary genres or periods.

3.3. Studies

3.3.1. Methodology

This section concentrates on the concrete research questions that have been addressed using text analysis tools. To characterise current practices, a number of representative or exemplary studies have been analysed. As a first step, an inventory was made of the practical studies which are discussed in the various contributions to the *Blackwell Companion to Digital Humanities* and the *Blackwell Companion to Digital Literary Studies*.[236] This initial list was extended by identifying all articles which discuss computer-based practical work, published either in *Literary and Linguistic Computing*, the *Digital Humanities Quarterly* or the *Journal of Digital Humanities* during the period in between 2009 and 2014. In the following sections, these studies have been clustered by considering the main literary phenomena they concentrate on.

3.3.2. Literary genres

Quantitative methods have often been used to study the stylistic differences between literary genres. Many of the studies in this category are based on analyses of word frequencies. Hugh Craig, for instance, has conducted a study of 25 Shakespeare plays, based of counts of the 12 most common words. The objective of

[236] *A Companion to Digital Humanities* (Malden, MA: Blackwell 2004) and Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Literary Studies* (Malden, MA: Blackwell 2007).

the study was to investigate if the groups that could be produced through statistical analyses of word frequencies correspond in some way to traditional divisions into comedies, tragedies and history plays. In Craig's study, a PCA revealed that the history plays indeed use a different set of words. Furthermore, there is a basic distinction between plays which contain a high frequency of the pronouns "you" and "I", on the one hand, and plays in which "of", "and" and "the" are commonly used. Craig argues that this reflects a division within the corpus between texts which contain "interactive dialogue" and plays which mostly contain "description and narration".[237]

Literary research on the differences between genres can alternatively be based on data about syntactic categories. This potential has been explored by some of the studies that have been conducted under the aegis of the Stanford Literary Lab. This lab was founded in 2010 by Matthew Jockers and Franco Moretti, and "discusses, designs, and pursues literary research of a digital and quantitative nature".[238] The first pamphlet of the Literary Lab addresses the question whether literary genres, such as the gothic novel or the Bildungsroman, can be recognised by computer algorithms. To investigate this issue, a range of experiments were carried out on British novels from the 19th century taken from the Chadwick-Healey collection.[239] The experiments focussed on two sets of data. The first set of data contained the LATS produced by DocuScope. The research was also based on information on the most frequent words. The data sets were analysed using principal component analysis and clustering technologies.[240] A central aim was to discover if certain combinations of LATs or frequent words are also characteristic of specific genres. The results were tested against genre assignments from existing bibliographies. Results suggested, however, that the techniques were best at recognising authorship, rather than genre. There were some notable differences between texts from different historical periods, but for genres that flourished during the same historical periods, the results were poor.[241]

The same topic was revisited in the fifth pamphlet of the literary lab. The pamphlet investigates the hypothesis that literary genres can be characterised through their use of specific grammatical constructions. It was also assumed that Gothic novels often contain fixed combinations of articles, nouns and prepositions, as in phrases like "the Castle of Otranto", or "the Rock of Glotzden".[242] To

---

[237] Hugh Craig, "Stylistic Analysis and Authorship Studies", in: *A Companion to Digital Humanities*, Oxford: Blackwell 2002, pp. 274-277.

[238] <http://litlab.stanford.edu/> (4 August 2013)

[239] <http://collections.chadwyck.co.uk/> (4 August 2013)

[240] More specifically, the "dist" and "hclust" functions available in the open-source "R" statistics application were used.

[241] Sarah Allison et al., *Quantitative Formalism: An Experiment*, (Stanford: Stanford Literary Lab 2011).

[242] Ryan Heuser & Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, (Stanford Literary Lab 2012), p. 2.

investigate these hypotheses, British novels were selected from the Chadwyck-Healey nineteenth-century database. From these novels, the sentences were isolated, and counts were created of particular types of sentences. Among other types, distinctions were made between sentences that consist of a single independent clause, sentences which contain an independent clause followed by a dependent clause, and sentences in which an independent clause is followed by a non-finite clause. In the case of sentences with multiple clauses, different types of conjunctions were also analysed. It was found, among other things, that Charles Dickens and Ann Radcliffe predominantly use sequencing conjunctions, in sentences that have a dependent clause preceded by an independent clause, and that Walter Scott mostly uses relative or "defining" dependent clauses. However, the findings were difficult to generalise into conclusive statements about correlations between grammatical constructions and literary genres.[243]

3.3.3. Literary characters

Stephen Ramsey has conducted a study which focused on Virginia Woolf's novel *The Waves*. The novel contains six related monologues, each spoken by distinct characters, who together narrate a related series of events. Ramsay has explored the differences between these six monologues on the basis of the term frequency-inverse document frequency (tf-idf) formula. In regular word lists, frequencies are usually distributed according to Zipf's Law, which states that there are normally small numbers of words that occur very frequently, and large numbers of hapax legomena, which are words that occur only once. The tf-idf formula assigns weights to the bare counts of the words, which are calculated by dividing the regular frequency of the type by the total number of texts that contain the type. Consequently, they have a higher value when words are infrequent. Ramsey demonstrates that the word lists that are generated in this way can indeed be used to disclose some of the central differences between the six protagonists. The list for the Australian character Louis, for instance, convincingly exposes a consciousness of his accent and of his nationality. Ramsey explains that statistical processing typically results in a paratext that can be interpreted and explained by the scholar, and that such resources may help to "to confirm or deny the 'serendipitous reading' of literary critics".[244]

Kyle Mahowald, at the Language Lab at MIT, has investigated occurrences of y- and th- pronouns as used by characters in Shakespeare's plays.[245] Making use of the Natural Language Processing toolkit in Python, a mechanism was developed to

---

[243] Ryan Heuser & Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, p. 10 and passim.

[244] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 14.

[245] K. Mahowald, "A Naive Bayes Classifier for Shakespeare's Second-Person Pronoun", in: *Literary and Linguistic Computing*, 27:1 (10 November 2011).

identify all occurrences of second person pronouns. In addition, a Naïve Bayes classifier was built to predict whether these pronouns were th- or y-pronouns, by making use of collocations. It was found that words such as “lordship”, “madam” and “sir” were most useful in classifying a pronoun as a y-pronoun, suggesting that these are mainly used to address a personal with a higher social status. Unsurprisingly, it was also found that th-pronouns were used mostly by characters with a higher status to address persons from lower social classes.

3.3.4. Date of creation

The studies that have been discussed so far are critically based on word segmentation and on analyses of frequency lists. Studies may also be based on random substrings. Richard Forsyth has investigated the differences between the early and the late poetry of W.B. Yeats using a technique which is called Monte Carlo Feature Finding. [246] The main aim of the study was to develop a method for dating texts. In this method, substrings are extracted in a manner that is fully opaque to the occurrence of word boundaries. Each substring is ranked according to its distinctiveness, which was measured on the basis of Chi-squared. The study suggested that short random substrings can indeed be used to classify texts either as an early or as a late poem. The methodology used in this particular study seems deeply remote from that of traditional literary studies. Substrings are, in most cases, entirely meaningless. While the algorithms enjoyed a degree of success in categorising poems, the patterns that were created are evidently difficult to interpret for literary critics. As such, the study fails to contribute to an understanding of how the early and the late poems differ precisely.

3.3.5. Authorship attribution

Word frequency lists have also been used successfully in studies that aim to suggest a probable author for texts whose authorship is disputed. Various studies have shown that patterns in the usage of vocabulary are strongly distinctive for individual authors, and that, as such, frequency lists can serve as an author's individual fingerprint. Authorship attribution studies generally make a distinction between lexical words and function words. The first term is used to refer to words such as nouns and verbs which usually carry most of the meaning in a sentence. Lexical words are often selected consciously, and the exact denotation or connotation usually varies strongly along with a word's context. Function words, by contrast, mostly have a relatively stable meaning. The term comprises words such as pronouns, articles and prepositions, which are assumed to be chosen unconsciously. Studies in the field of authorship attribution predominantly focus

[246] Richard S Forsyth, “Stylochronometry with Substrings, or: A Poet Young and Old”, in: *Literary and Linguistic Computing*, 14:4 (1999).

on function words. They make use of a "base strata of language where imitation or deliberate variation can be ruled out".[247]

Authorship attribution studies usually demand a meticulous preparation of the source materials, since such investigations largely exploit the distributions of function words which are "especially resistant to intentional authorial manipulation".[248] Burrows notes that, in the case of older texts, it can be useful to standardise the spelling or to expand contracted forms.[249] In addition, certain homographic words can also be tagged so that the different grammatical uses of words can be distinguished. Hoover has found that the accuracy of attribution tests improves when proper nouns, inflected words, personal pronouns are removed from the corpus. Hoover also proposes a removal of all dialogue from novels, but since dialogue is not always rendered typographically distinct, this often requires a degree of interpretation.

A large number of authorship attribution studies have made use of the delta method, which was developed originally by John Burrows. The method assumes a corpus which contains a work whose author is unknown, together with works by a number of potential authors. The delta value of a work, or a group of works, can be calculated by considering the difference between "the z-scores for a set of word-variables in a given text group and the z-scores for the same set of word-variables in a target set".[250] The delta value is the mean of the absolute values of these differences. Importantly, delta ignores the difference between positive and negative values. The central assumption in Burrow's method is that the probable author of the unassigned work can be found by comparing the delta value for this individual work to the values for the potential authors. If the delta is low, there is a higher probability that the works are by the same author. In his article *Never Say Always Again: Reflections on the Numbers Game*, Burrows discusses a number of exemplary authorship attribution studies which are based on delta. The first of these concerns the novel *St. Ives*, which was begun by Robert Louis Stevenson, and which was completed by Arthur Quiller Couch after Stevenson's death in 1894. For the study, Burrows took 72,000-word samples from the novels and stories of Stevenson and Quiller Couch, together with 12,000-word samples from authors who were active during roughly the same period. The tests pointed to Stevenson as

[247] Hugh Craig, "Stylistic Analysis and Authorship Studies", p. 273.

[248] David Hoover, "Word Frequency, Statistical Stylistics and Authorship Attribution", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009, p. 35.

[249] John Burrows, "Textual Analysis", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, Oxford: Blackwell 2002, p. 269.

[250] J. Burrows, "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing*, 17:3 (1 September 2002), p. 271.

the author of the first 29 chapters, and to Quiller Couch for the authorship or the remaining 11 chapters, which is in line with the available historical evidence.[251]

3.3.6. Themes

Analyses of the words that are used within a corpus can also be applied to find the words associated with specific themes or specific topics. Martha Nell Smith has explored the use of erotic language in the correspondence of Emily Dickinson, on the basis of a corpus of about 200 letters that were exchanged between the poet Emily Dickinson and her sister-in-law Susan Huntington Dickinson. As part of the study, an application was built in which scholars could classify the degree of eroticism, on a scale of 0 to 5. The letters that were rated manually formed the training material for the classifier. The application attempted to classify the remaining texts, using a method based on Naieve Bayesian logic. The application also indicated the words that were found to be associated with eroticism. One of the surprising outcomes was that the word "mine" emerged as a marker of eroticism. According to Smith, certain words only assume an erotic subtext because of their co-occurrence with other words, and computers are better at exposing such co-occurrences because human readers "may unselfconsciously divide epistolary subjects within the same letter [...] into completely separate categories".[252] The complete disregard for rhetorical structure may help to detect connections which may previously have escaped scholars.

Kao and Jurafsky have investigated the literary quality of poems, by comparing poetry written by skilled professional poets to texts written by amateur poets.[253] Assuming that the first class of poems is of a higher literary quality than the poetry in the latter class, Kao and Jurafsky examined the differences between these two sets of poems through a quantification of some of the linguistic and semantic features. The latter data were obtained by making use of HGI and LIWC. The counts obtained via these semantic taggers were normalised for the length of the poem. To investigate the differences between professional and amateur poetry, a logistic regression model for all 16 metrics was implemented in the R package. It was found that, while professional poets are significantly less likely to use words that explicitly refer to negative emotions than amateur poets, the usage of words with negative connotations was found to be roughly the same. This suggests that poets mainly evoke sentiments through connotations rather than through direct

[251] John Burrows, "Never Say Always Again: Reflections on the Numbers Game".

[252] Martha Nell Smith et al., ""Undiscovered Public Knowledge": Mining for Patterns of Erotic Language in Emily Dickinson's Correspondence with Susan Huntington (Gilbert) Dickinson", in *Digital Humanities*, (2006), pp. 252–255, p. 254.

[253] Justine Kao & Dan Jurafsky, "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry", in *NAACL Workshop on Computational Linguistics for Literature*, (2012), pp. 8–7.

denotation. In addition, the results also indicated that professional poets tend to use concrete imagery rather than references to abstract or general concepts.

Archer, Culpepper and Rayson have used a related method to investigate "key domains" in Shakespeare's drama.[254] The authors have investigated the 'aboutness' of six comedies and tragedies using USAS. The semantic categories that were found enabled the authors to investigate the images that served as the tenors in metaphors that were used in the love comedies and in the love tragedies. The differences between the two sets of poems were compared by calculating the log-likelihood from the frequencies of the semantic tags that were identified. Through this procedure, the most overused and the most underused semantic fields could be identified for both groups of plays. It was found, among other things, that the categories 'liking' and 'intimate/sexual relationships' was underused in the love tragedies, in comparison to the love comedies. The tragedies had a higher occurrence of tags such as 'war' and 'religion and the supernatural'.

Lisa Rhody has used Topic Modelling to investigate figurative language in a genre of poetry called Ekphrasis, which entails "poems written to, for, or about the visual arts".[255] Rhody used LDA to identify topics in ca. 4,500 poems, and encountered a large number of difficulties. An obvious complication was that, while Topic Modelling assumes that there is a close connection between the words that are used and the topics they refer to, figurative language obviously uses word senses in ways that differ widely from their conventional meanings. Furthermore, ekphrastic poetry typically discusses works of art which depict certain phenomena. The phenomena that are shown in paintings or in drawings, however, do not necessarily constitute the central topics of the poem. A number of topics veered felicitously towards specific images and specific themes, but MALLET also returned various lists of words which seemed to have little in common. Such "semantically opaque" topics are frequently of particular interest for the literary critic, however, as they may reveal particular forms of discourse. They can also indicate words that are typical for specific genres or for specific themes. Rhody's study has led to an improved understanding of the words that are commonly used in elegiac poetry, for instance. Furthermore, lists of words which seem to have little in common may stimulate scholars to revisit the poems in which these words occur, and such more directed forms of close reading can show that these texts share particular thematic concerns. Rhody argues that working with MALLET can help scholars to make new discoveries, "not because topic modelling works

[254] Dawn Archer, Jonathan Culpeper & Paul Rayson, "Love – "a Familiar or a Devil"? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009.

[255] Lisa M. Rhody, "Topic Modeling and Figurative Language", in: *Journal of Digital Humanities*, 2:1 (2012).

perfectly, but because poetry causes it to fail in ways that are potentially productive for literary scholars".[256]

3.3.7. Lexical repetitions

When a text has been tokenised into words, it is also possible to study repetitions of words or groups of words in a text. In a pioneering study, Tanya Clement has applied several algorithms for the exploration of repetitive patterns in the novel *The Making of America* by Gertrude Stein, a postmodern work which in itself is "almost impossible to read [...] in a traditional, linear manner".[257] Software has been developed which divided the text into so-called n-grams, or patterns of co-occurring words. The study focused, more specifically, on series of three consecutive words. It was shown that the exact same trigram re-occurred in different sections of the novel. Moreover, in a number of passages, verbatim repetitions of even longer units were found. Such repetitions are difficult to identify without a computer, since full-text searches for specific repeated phrases assumes a "pre-knowledge that they exist—a nontrivial feat in the midst of the more pervasive and shorter repetitions that make up each section".[258] As part of the study, a number of visualisation tools have also been devised to indicate how repetitive patterns are distributed over the various sections of the novel. Clement's analyses suggested that the placement of repetitions indeed follows a highly regular pattern. While many studies in the field of literary informatics study vocabulary under the bag of words model, in which the original context of words is lost, this approach is evidently unsuitable for studies that focus on the structural repetition of words and phrases, as information about the original location of the words is generally relevant for studying the nature and the distribution of such echoes. In such studies, a bag of words model is inappropriate.

3.3.8. Rhyme and meter

Various scholars have used the computer to support of prosodic analyses. Malcolm Hayward, an English scholar at Indiana University of Pennsylvania, has conducted an exemplary study which focussed on 1000 lines of poetry, consisting, more specifically, of ten 100-line samples from ten different authors, including John Donne, Alexander Pope, John Keats and Robert Browning. All lines in the corpus were iambic pentameters. Hayward's aim was to "provide a quantitative basis for

[256] Lisa M. Rhody, "Topic Modeling and Figurative Language", n.pag.
[257] T. E. Clement, ""A Thing Not Beginning and Not Ending": Using Digital Tools to Distant-Read Gertrude Stein's The Making of Americans", in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), p. 361.
[258] Ibid., p. 363.

comparisons between poets",[259] and to investigate the metrical variations in these poetic lines, using a connectionist model of poetic meter. The connectionist model is based on the idea that each syllable in the line "is connected to five other units, representing possible inputs towards stress from intonation, lexical features, prosody, syntax, and interpretation". Hayward has manually assigned scores for each of these dimensions. At the lexical level, for example, primary stress was marked with a '2', secondary stress with a '1', and all unstressed syllables received '0'. The scores were analysed using statistical software, and this resulted in "a measurement of the potential activation of metrical stress for each of the ten positions for that particular line of poetry". Next, multivariate analyses were performed for the stress of each individual syllable, and these revealed significant differences among all ten poets represented in the study. From these findings, the conclusion was drawn that poets indeed display a highly idiosyncratic behaviour with respect to metrical variation.

Experiments with automated phonetic transcriptions have been conducted by, amongst others, David Kaplan. Kaplan was interested in the question if differences and similarities between poems can also be displayed visually. To explore this question, Kaplan developed a series of algorithms for the quantification of specific features of texts written by Northern-American poets.[260] Kaplan has also developed a number of algorithms that could use these transcriptions, together with data on syntactic categories produced by a POS Tagger, to produce a total of 84 different metric values for each poem. In the study, data were produced about occurrences of alliteration, assonance, consonance, perfect rhyme, slant rhyme and half rhyme, amongst other aspects. Automating the scansion is complicated by the fact that the placement of stresses depends, to a large degree, upon the meaning of the line. Using Principal Component Analysis and metric CMDS, these multiple values were visualised in a two-dimensional plot. The software developed "showed an ability to distinguish poetry texts based on a combination of salient features not traditionally used in computational prose text analysis but traditionally relied upon for poetry analysis".[261]

3.3.9. Allusions

Data on grammatical categories and on lemmas can also be applied usefully in explorations of literary allusions. Walter Crane explains that allusion can be viewed as a reference of one text fragment to another text fragment, and that they can be

[259] Malcolm Hayward, "Analysis of a Corpus of Poetry by a Connectionist Model of Poetic Meter", in: *Poetics*, 24:1 (July 1996), n.pag.
[260] David Kaplan & D.M. Blei, "A Computational Approach to Style in American Poetry".
[261] David Maxwell Kaplan, *Computational Analysis and Visualized Comparison of Style in American Poetry* (Princeton University 2006), p. 31.

either direct or indirect.[262] Direct allusions, firstly, are verbatim repetitions. The difference with quotations is that imitative textual allusions also invoke the context of the original, and that a knowledge of the reference is needed to interpret the allusion. Indirect allusions are essentially reworked versions of the original text, and they are, for this reason, more difficult to recognise algorithmically. Crane distinguishes a number of different types. Two fragments may contain identical words in a different order. Two passages may also share lemmas. Additionally, there may be a syntactical identity, a metrical identity and a semantic identity. Coffee et al. have noted that while computers may be used to identify such textual parallels, the automated detection of allusions is complicated by the fact that textual parallels are not always of literary significance. To distinguish meaningful allusions from other forms of parallels, the researchers have devised a model using data supplied by human critics. Parallels based on an "expanded feature set including bi-gram frequency, frequency of individual words, character-level n-grams and edit distances" were analysed through "support vector machines" and "random forests", and these experiments resulted in a number of statistical rules to identify the characteristics of meaningful allusions.[263] The algorithms have also been used to explore allusions to Vergil's *Aeneid* in the first book of Lucan's *Civil War*. The tool produced over 2,500 textual parallels, but many of these were found to be irrelevant. Nevertheless, the results that have been generated by Tesserae enables the developers to train and to further refine the algorithms.[264]

3.4. Discussion

As was discussed, computers can produce data about a broad range of textual aspects. On the basis of the spaces that are used in between words, computer applications can produce counts of the total number of words and of the total number of unique words. Tools in the field of NLP can additionally be used to produce data about the syntactic or grammatical categories of words, their lemmatised forms, their pronunciation, and their signification. Once they have been collected, these multifarious data can be analysed and visualised in a myriad of ways. In recent decades, quantitative approaches have been applied to study the formal aspects of literary genres, characters, themes, allusions and literary themes, among many other topics.

Notably, many of the studies which have been surveyed in this chapter have focused primarily on the further development and refinement of the methodology,

[262] Gregory Crane & David Bamman, "The Logic and Discovery of Textual Allusion", in: *ACL Language Technology for Cultural Heritage*, (2008).

[263] Neil Coffee et al., "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques", in *Digital Humanities 2013*, (Nebraska–Lincoln: 2013), n.pag.

[264] N. Coffee et al., "The Tesserae Project: Intertextual Analysis of Latin Poetry", in: *Literary and Linguistic Computing*, 28:2 (20 July 2012).

rather than on the creation of new knowledge about literary texts. In *Testing Burrows's Delta*, for instance, David Hoover assesses the effectiveness and the accuracy of the delta method, and, using a sample corpus in which authors are all known already, demonstrates that the accuracy increases when a larger number of words are taken into account, or when personal pronouns are removed.[265] In a comparable study, Eder aimed to quantify the impact of textual errors on the accuracy of authorship attribution methods. The researchers had introduced varying numbers of typing errors in text corpora, in order to determine which level of noise would be acceptable.[266] The objective of the studies that were carried out for the first pamphlet of the Stanford Literary Lab was similarly to develop a procedure that can be used to create the exact same clusters and classifications that have been produced earlier by human scholars.[267] The assumption is that when the algorithms work correctly for a sample corpus of a modest size, the same rule-based approach can also be followed to classify texts which had previously been neglected.

As the current methodology is still of a probationary and experimental nature, studies which aim to calibrate or to meliorate the toolset are very beneficial. The ultimate aim of literary informatics research, however, is to concoct and to implement analytic methods which can genuinely advance the emanation of new ideas and new insights. Scholars ought to explore the ends to which these instruments can be put, and they need to evaluate the relevance or the value of these new methods, by enlisting these in the service of broader humanistic questions. Algorithmic criticism initially converts works of literature into numbers, but, ultimately, these numbers need to be converted in turn into qualitative or interpretative statements which can challenge, confirm or enrich our understanding of the texts that have been quantified.

[265] D. L. Hoover, "Testing Burrows's Delta", in: *Literary and Linguistic Computing*, 19:4 (1 November 2004).
[266] M. Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Literary and Linguistic Computing*, (14 November 2013).
[267] Sarah Allison et al., *Quantitative Formalism: An Experiment*.