Exploring Cities in Crime: Significant Concordance and Co-occurrence in Quantitative Literary Analysis

Janneke Rauscher[1] **Leonard Swiezinski**[2] **Martin Riedl**[2] **Chris Biemann**[2]
(1) Johann Wolfgang Goethe University
Grüneburgplatz 1, 60323 Frankfurt am Main, Germany
(2) FG Language Technology, Dept. of Computer Science
Technische Universität Darmstadt, 64289 Darmstadt, Germany
j.rauscher@em.uni-frankfurt.de, floppy35@web.de,
{riedl,biem}@cs.tu-darmstadt.de

Abstract

We present CoocViewer, a graphical analysis tool for the purpose of quantitative literary analysis, and demonstrate its use on a corpus of crime novels. The tool displays words, their significant co-occurrences, and contains a new visualization for significant concordances. Contexts of words and co-occurrences can be displayed. After reviewing previous research and current challenges in the newly emerging field of quantitative literary research, we demonstrate how CoocViewer allows comparative research on literary corpora in a project-specific study, and how we can confirm or enhance our hypotheses through quantitative literary analysis.

1 Introduction

Recent years have seen a surge in Digital Humanities research. This area, touching on both the fields of computer science and the humanities, is concerned with making data from the humanities analysable by digitalisation. For this, computational tools such as search, visual analytics, text mining, statistics and natural language processing aid the humanities researcher. On the one hand, software permits processing a larger set of data in order to assess traditional research questions. On the other hand, this gives rise to a transformation of the way research is conducted in the humanities: the possibility of analyzing a much larger amount of data – yet in a quantitative fashion with all its necessary aggregation – opens the path to new research questions, and different methodologies for attaining them.

Although the number of research projects in Digital Humanities is increasing at fast pace, we still observe a gap between the traditional humanities scholars on the one side, and computer scientists on the other. While computer science excels in crunching numbers and providing automated processing for large amounts of data, it is hard for the computer scientist to imagine what research questions form the discourse in the humanities. In contrast to this, humanities scholars have a hard time imagining the possibilities and limitations of computer technology, how automatically generated results ought to be interpreted, and how to operationalize automatic processing in a way that its unavoidable imperfections are more than compensated by the sheer size of analysable material.

This paper resulted from a successful cooperation between a natural language processing (NLP) group and a literary researcher in the field of Digital Humanities. We present the CoocViewer analysis tool for literary and other corpora, which supports new angles in literary research through quantitative analysis.

In the Section 2, we describe the CoocViewer tool and review the landscape of previously available tools for our purpose. As a unique characteristic, CoocViewer contains a visualisation of significant concordances, which is especially useful for target terms of high frequency. In Section 3, we map the landscape of previous and current quantitative research in literary analysis, which is still an emerging and somewhat controversial sub-discipline. A use-case for the tool in the context of a specific project is laid out in Section 4, where a few examples illus-

Proceedings of the Second Workshop on Computational Linguistics for Literature, pages 61–71,
Atlanta, Georgia, June 14, 2013. ©2013 Association for Computational Linguistics

trate how CoocViewer is used to confirm and generate hypotheses in literary analysis. Section 5 concludes and provides an outlook to further needs in tool support for quantative literary research.

2 CoocViewer - a Visual Corpus Browser

This section describes our CoocViewer visual corpus browsing tool. After shortly outlining necessary pre-processing steps, we illustrate and motivate the functionality of the graphical user interface. The tool was specifically designed to aid researchers from the humanities that do not have a background in computational linguistics.

2.1 Related Work

Whereas there exist a number of tools for visualizing co-occurrences, there is, to the best of our knowledge, no tool to visualize positional co-occurrences, or as we also call them, significant concordances. In (Widdows et al., 2002) tools are presented that visualize meanings of nouns as vector space representation, using LSA (Deerwester et al., 1990) and graph models using co-occurrences. There is also a range of text-based tools, without any quantitative statistics, e.g. Textpresso (Müller et al., 2004), PhraseNet[1] and Walden[2]. For searching words in context, Luhn (1960) introduced KWIC (Key Word in Context) which allows us to search for concordances and is also used in several corpus linguistic tools e.g. (Culy and Lyding, 2011), BNCWeb[3], Sketch Engine (Kilgarriff et al., 2004), Corpus Workbench[4] and MonoConc (Barlow, 1999). Although several tools for co-occurrences visualization exist (see e.g. co-occurrences for over 200 languages at LCC[5]), they often have different aims, and e.g. do not deliver the functionality to filter on different part-of-speech tags.

2.2 Corpus Preprocessing

To make a natural language corpus accessible in the tool, a number of preprocessing steps have to be carried out for producing the contents of CoocViewer's database. These steps consist of a fairly standard natural language processing pipeline, which we describe shortly.

After tokenizing, part-of-speech tagging (Schmid, 1994) and indexing the input data by document, sentence and paragraph within the document, we compute signficant sentence-wide and paragraph-wide co-occurrences, using the tinyCC[6] tool. Here, the log-likelihood test (Dunning, 1993) is employed to determine the significance $sig(A, B)$ of the co-occurrence of two tokens A and B. To support the significant concordance view (described in the next section), we have extended the tool to also produce *positional* significant co-occurrences, where $sig(A, B, offset)$ is computed by the log-likelihood significance of the co-occurrence of A and B in a token-distance of $offset$. Since the significance measure requires the single frequencies of A and B, as well as their joint frequency *per positional offset* in this setup, this adds considerable overhead during preprocessing. To our knowledge, we are the first to extend the notion of positional co-occurrence beyond direct neighbors, cf. (Richter et al., 2006). We apply a sigificance threshold of 3.84[7] and a frequency threshold of 2 to only keep 'interesting' pairs. The outcome of preprocessing is stored in a MySQL database schema similar to the one used by LCC (Biemann et al., 2007). We store sentence- and paragraph-wide co-occurrences and positional co-occurrences in separate database tables, and use one database per corpus. The database tables are indexed accordingly to optimize the queries issued by the CoocViewer tool. Additionally, we map the part-of-speeches to E (proper names), N (proper nouns), A (adjectives), V (verbs), R (all other part-of-speech tags) for an uniform representation for different languages.

2.3 Graphical User Interface

The graphical user interface (UI) is built with common web technologies, such as HTML, CSS and JavaScript. The UI communicates via AJAX with a backend, which utilizes PHP and a MySQL

[1] http://www-958.ibm.com/software/data/cognos/manyeyes/page/Phrase_Net.html

[2] http://infomotions.com/sandbox/network-diagrams/bin/walden/

[3] http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php

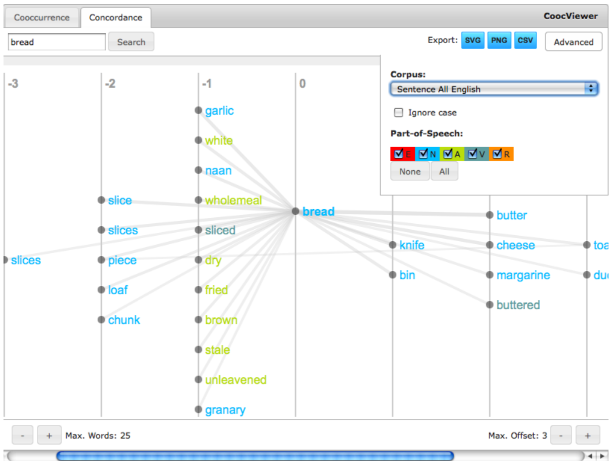
[4] http://cwb.sourceforge.net

[5] http://corpora.uni-leipzig.de/

[6] http://wortschatz.uni-leipzig.de/~cbiemann/software/TinyCC2.html, (Richter et al., 2006)

[7] corresponding to 5% error probability

database. This makes the approach flexible regarding the platform. It can run on client computers using XAMP[8], a portable package of various Web technologies, including an Apache web server and a MySQL server. Alternatively, the tool can operate as a client-server application over a network. In particular, we want to highlight the JavaScript data visualization framework D3 (Bostock et al., 2011), which was used to layout and draw the graphs. We deliberately designed the tool to match the requirements of literary researchers, who are at times overwhelmed by general-purpose visualisation tools such as e.g. Gephi[9]. The UI is split into three parts: At the top a menu bar, including a search input field and search options, a graph drawing panel and display options at the bottom of the page.



bread

Corpus: Sentence All English

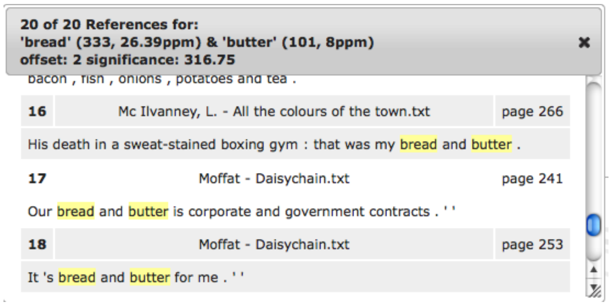
☐ Ignore case

Max. Words: 25

Max. Offset: 3

Figure 1: Screenshot of the Coocviewer application using the concordance view.

The menu bar allows switching between co-occurrence and concordance views (see Figure 1). The search field supports wildcards and type-ahead autocompletion, immediately displaying which words exist in the corpus and match the current input. Additionally, there are functionalities to export the shown graph as SVG or PNG image, or as plain text, containing all relations, including their frequencies and significance scores. Within the advanced configuration windows (shown on the right side) one can select different corpora, enable case sensitive/insensitive searches or filter words according their part-of-speech tags (as described in Section 2.2). The graph drawing panel visualizes the queried term and its significant co-occurrences resp. concordances, significancy being visualized by the thickness of the lines. In Figure 1, showing the concordances for *bread*, we can directly see words that occur often with *bread* in the context: E.g. *bread* is often used in combination with *butter, cheese, margarine* (offset +2), but also the kind of different breads is described by the adjectives at offset -1. For the same information, using the normal KWIC view, one has to count the words with different offset by hand to find properties for the term *bread*. At the bottom, the maximal number of words shown in the graph can be specified. For the concordances display there is an additional option to specify the maximal offset. The original text (with provenance information) containing the nodes (words) or edges (word pairs) shown in the graph can be retrieved by either clicking on a word itself or on the edge connecting two words, in a new window (see Figure 2) within the application. This window also provides information about the frequencies of single words as well as their co-occurrence, and also displays relative single word frequencies in parts-per-million (ppm) to enable comparisons between corpora of different sizes. Words in focus are highlighted and the contents of this window can also be exported as plain text.



20 of 20 References for:
'bread' (333, 26.39ppm) & 'butter' (101, 8ppm)
offset: 2 significance: 316.75

bacon , fish , onions , potatoes and tea .

16	Mc Ilvanney, L. - All the colours of the town.txt	page 266

His death in a sweat-stained boxing gym : that was my bread and butter .

17	Moffat - Daisychain.txt	page 241

Our bread and butter is corporate and government contracts . ' '

18	Moffat - Daisychain.txt	page 253

It 's bread and butter for me . ' '

Figure 2: Occurrences of a significant concordance

3 Quantitative Literary Research

Quantitative research in literary analysis, although being conducted and discussed since at least the 1960s, (Hoover, 2008), is still far from being a clear field of research with a verified and acknowledged methodology. Studies in this field vary widely with respect to scope, methods applied and theoretical

[8] `http://www.apachefriends.org/en/index.html`

[9] `https://gephi.org/`

background. Until now, only the most basic definition can be given that applies to these approaches: Quantitative research in literary analysis is generally concerned with the application of methods from corpus linguistics (and statistics) to the field of literature to investigate and quantify general grammatical and lexical features of texts.

Most studies applying such methods to literary analysis are carried out in the field of stylistics, building a relatively new research area of corpus stylistics, also called stylometry (Mahlberg, 2007; Hoover, 2008; Biber, 2011). The quantitative exploration of stylistic features and patterns is used for authorship attribution, e.g. (Burrows, 1992; Burrows, 2007; Craig, 2004; Hoover, 2001; Hoover, 2002), exploring the specificity of one author's style, e.g. (Burrows, 1987; Hori, 2004; Fischer-Starcke, 2010; Mahlberg, 2012) or one certain text, often compared to other texts of the same author or period, e.g. (Craig, 1999; McKenna and Antonia, 2001; Stubbs, 2005; Clement, 2008; Fischer-Starcke, 2009). Some studies focus on content-related questions such as the analysis of plot or characterization and the exploration of relations between and role of different characters, e.g. (Mahlberg, 2007; Culpeper, 2002; Culpeper, 2009), developing new ways of exploring these literary features, e.g. via the application of social network analysis (Elson et al., 2010; Moretti, 2011; Agarwal et al., 2012). Besides this area, there are numerous other approaches, like the attempt to investigate the phenomenon of "literary creativity" (Hoey, 2007) or ways for automatic recognition of literary genres (Allison et al., 2011).

Major methodological approaches of this field are, according to Biber (2011), Mahlberg (2007) and Hoover (2008), the study of keywords and word-frequencies, co-occurrences, lexical clusters (also called bundles or n-grams) and collocational as well as concordance analysis. Additionally, the need for cross-investigating and comparing the results with other corpora (be it a general corpus of one language or other small, purpose-built corpora) is emphasized to discuss the uniqueness of the results.

But while especially the studies of Moretti (2000; 2007; 2009), taking a quantitative approach of "distant reading" on questions of literary history and the evolution of literary genres, are often received as groundbreaking for the case, and despite the rising interest in this field of research in the last decades, there still is much reluctance towards the implementation of such methods. The general arguments raised frequently from the point of view of 'classical' literary analysis against a quantitative or computational approach can be grouped around four central points: The uniqueness of each literary text that quantitative analysis seems to underscore when treating the texts just as examples of style or period, focussing on very general patterns; the emphasize of technology and the relatively high threshold that the application, analysis and interpretation of the generated data contains (Potter, 1988); and the general notion that meaning in literary texts is highly context-related and context-dependent in different ways (Hoover, 2008). Last but not least there is what can be called the "so-what-argument": Quantitative methods tend to produce sparse significant new information compared with the classical approach of close reading, generating insights and interpretations that could as well be reached by simply reading the book (Mahlberg, 2007; Rommel, 2008). But the possibilities and advantages of corpus linguistics come to the foreground especially if one is not interested in aspects of uniqueness or particularity but in commonalities and differences between large amounts of literary texts too many to be read and compared in the classical way. This especially holds when it comes to questions of topics, themes, discourse analysis and the semantisation of certain words.

4 Empirical Analysis

This section describes a few exemplary analysis which we carried out within our ongoing project "At the crime scene: The intrinsic logic of cities in contemporary crime novels". Settled between the disciplines of sociology and literature, the project is embedded in the urban sociological research area of the 'Eigenlogik' (intrinsic logic) of cities (Berking, 2012; Löw, 2012; Löw, forthcoming). The basic hypothesis is that there is no such thing as 'the' city or 'the' urban experience in general, but that every city forms its own contexts and complexes of meaning, the unquestioned and often subconsciously operating knowledge of how things are done, respectively making sense of the city. To put it another way, we want to find out if and in what way the respective city makes a difference and is forming distinctive structures of thought, action and feeling. This is explored simultaneously in four different projects investigating different fields (economic practices, city marketing, problem discourses and literary field and texts) in four different cities that are compared with each other (Birmingham (UK), Glasgow, Frankfurt on the Main and Dortmund). If the hypothesis is right, the four different investigated fields should have more in common within one city and across the fields than within one field across different cities.

Our subproject is mainly concerned with the literary and cultural imagination and representation of the cities in question. One crucial challenge is the exploration, analysis and comparison of 240 contemporary crime novels, each of them set in one of the cities under examination. The aim of this explorative study is to analyze the possibility and characteristics of city-specific structures within the realm of literary representations of cities.

Dealing with such comparably large amounts of literary texts, a tool was needed that facilitated us (laypeople in the field of corpus linguistics) to explore the city-specific content and structures within these corpora, enabling a connection of qualitative close reading and quantitative methods. Visualization was a major concern, apparently lowering the resistance of the literary research community towards charts and numbers and making the results readable and interpretable without having much expertise in corpus linguistics. Moreover, the option of generating significant concordances instead of simple concordance lines (as e.g. with KWIC) is very promising: Confronted with very high word frequencies for some of our search terms, e.g. more than 2200 occurrences of "Frankfurt" in our Frankfurt corpus, completely manual analysis turned out to be painstaking and very time-consuming. Automated or manual reduction of the number of lines according to standard practices, as e.g. suggested by Tribble (2010), is not possible without potential loss of information. CoocViewer enables a sophisticated and automated analysis with concentration on statistically significant findings through clustering co-occurring words according to their statistical significance in concordance lines. Additionally, the positionality of these re-occurring co-occurrences in relation to the search term (with a maximum range from -10 to +10 around the node) gives a clear and immediate picture of patterns of usage within a corpus. Via exploring the references of the results we are still able to take account of the context-specificity of literary texts, as well as distinguishing author-specific results from those distributed more equally across a corpus.

City	lang.	#novels	#tokens	#sent.	#para.
Birmingham	engl.	41	4.8M	336K	142K
Glasgow	engl.	61	7.7M	496K	222K
Dortmund	ger.	59	5.0M	361K	127K
Frankfurt	ger.	79	8.0M	546K	230K

Table 1: Quantitative characteristics of our corpora

After describing the corpus resources, we conduct two exemplary analysis to show how the quantitative tool as described in Section 2 can be used to aid complex qualitative research interests in the humanities through supporting the exploration and comparison of large corpora (Sect. 4.2), as well as investigating and comparing the semantization and semantic preference of words (Sect. 4.3). The discussion of results shows how CoocViewer can support hypothesis building and testing on a quantitative basis, linking qualitative and quantitative approaches.

4.1 Corpus

The selection of the crime novels was based on three criteria: contemporariness (written and published within the past 30 years until 2010), the city in question (should play a major role resp. be used as major setting), and genre (crime fiction in any variety). In a first step, the 240 novels (gathered as paperback-editions) had to be scanned and digitalized[10]. Metadata was removed and the remaining texts were preprocessed as described in Section 2.2. The novels were compiled in different corpora according to the city they are set in, and the database underlying them (sentence or paragraph). Table 1 provides an overview of the quantitative characteristics of the four city-specific corpora we discuss here.

[10] We used ABBY FineReader 10 professional for optical character recognition, which generated tolerable but not perfect results, making extensive proof reading and corrections necessary.

4.2 Analysis 1: Exploring the Use of the City's Name

The occurrence of the name of a city in crime novels can serve different purposes and functions in the text. It can be used, for example, to simply 'place' the plot (instead of or additionally to describing the setting in further detail) or to indicate the direction of movement of figures ("they drove to Glasgow"). Often it is surrounded by information about city-specific aspects, e.g. of history or materiality. Searching for the respective proper names of the cities in the four corpora therefore seems to be a promising start to explore the possibility of city-specific structures of meaning in literary representation. If the 'Eigenlogik'-hypothesis is right, not only the content that is associated with the name (what would generally be expected) but also its frequent usages and functions (as pointer or marker, as starting point for further explanations of city life, etc.) should differ systematically across cities.

A first close reading of some exemplary crime novels already suggested that this could be the case. To check this qualitatively derrived impression we conducted CoocViewer searches for the top-15 significant co-occurrences across all parts of speech for each proper name in the respective corpus on sentence level (see Figure 3 for the cases of Glasgow and Frankfurt). To interpret and compare these findings, we additionally looked at the significant concordances (with the same search parameters and an offset from the node of -3/+3), which helps to analyze and refine the findings in more depth. In the following, we discuss, compare and interpret the results with respect to our overriding project-hypothesis to verify or falsify some of our qualitative first impressions quantitatively.

The corpora indeed tend not only to vary significantly with respect to the sheer frequency of the usage of the proper name (with relative frequencies ranging from Glasgow (324ppm) and Frankfurt (286ppm) to Dortmund (187ppm) and Birmingham (154ppm)), but also in the usages and functions that the naming fulfills. The graphs reveal not only differing co-occurrences, but also differing proportions of co-occurring word classes, each city revealing its own distinct pattern.

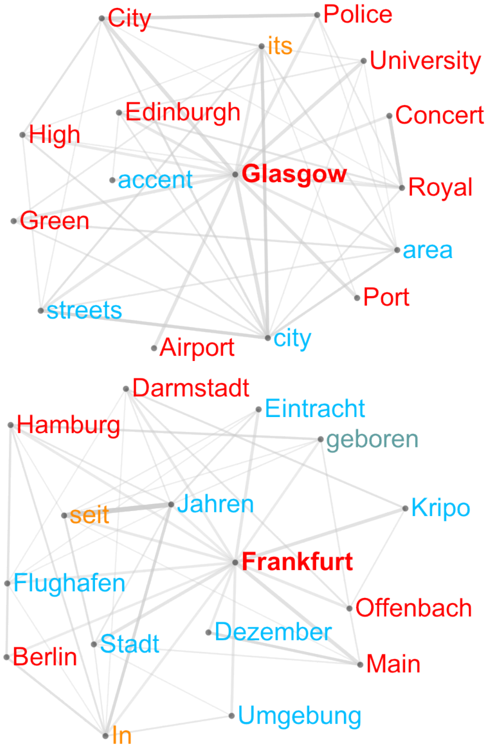


Figure 3: Significant co-occurrences of "Glasgow" (upper) and "Frankfurt" (lower) in their respective corpora

Especially the English cities tend to co-occur with proper names and common nouns (ten proper names, four common nouns in the case for Glasgow, eight names and six nouns for Birmingham).[11] For Glasgow, these comprise parts of the inventory of the city (with "City" (sig. of 695.57) as either part of the name or city-specific institution ("City of Glasgow", "City Orchestra") or to refer to different crime-genre specific institutions (as the "City of Glasgow Police" or "Glasgow City Mortuary")), the "University" (sig. of 380.42), or the park "Glasgow Green" (233.46). There is also the name of another city, the Scottish capital (and rival city) Edinburgh. As the statistical concordances reveal quickly, the "Port" (350.88) is, despite Glasgow's history as shipbuilding capital, not used to refer to the cities industrial past. Instead, as can bee seen from its positioning on -1 directly left to the node, it refers to the small nearby town Port Glasgow (see

[11] The noun "accent" which both English cities names co-occur with (and for which no equivalent term can be found on the German side) can be explained by a different lexicalization of the concept, which is realized through derivation in German.

Fig. 4). The co-occurrence of "Royal" and Glasgow (being not a royal city) can also be easily explained via the concordance view, showing that this is mainly due to the "Royal Concert Hall" (forming a strong triangle on positions +1, +2 and +3 from the node). Besides these instances of places and institutions within and around the city, especially the connection to the pronoun "its" (82 instances with a sig. of 144) is interesting. None of the other cities shows a top-significant co-occurrence with a comparable pronoun. A look at the corresponding references in the corpus shows that it is mainly used in statements about the quality or speciality of certain aspects of the city (indicated on graphic level through the connections between "its" and "city" or "area") and in personifications (e.g. "Glasgow could still reach out with its persistent demands"). This implies that the literary device of anthropomorphization of the city (in direct connection with the proper name) occurs more often within Glasgow-novels than within those of the other cities, and that there are many explicit statements about "how this city (or a special part of it) is", showing a tendency to explain the city. Furthermore, the exploration of the different references indicates a relatively 'coherent corpus' (and, therefore, relatively stable representation) with recurring instances across many authors.

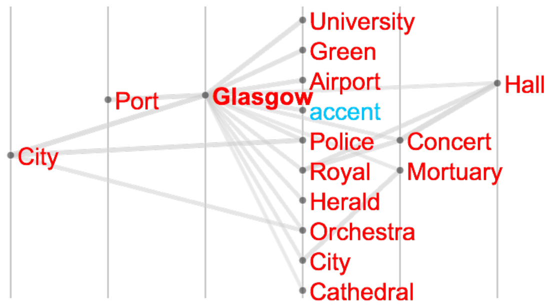


Figure 4: Significant concordances of "Glasgow" in Glasgow corpus

In contrast to this, Birmingham's co-occurring proper names mainly refer to (fictive) names of newspapers (the Birmingham "Sentinel", "Post" and "News"). The inventory of the city is not very prominently represented, with "University" (sig. of 152.52) and "Airport" (80.63) as the only instances. Furthermore, the University tends to be represented as region-, not city-specific (with a stronger connection between "University" and "Midlands" (sig. of 200.49) than between both words to the city itself ("Midlands" co-occurring with a sig. of 68.68)). The rest of the proper names relates to not further specified parts of the city ("East" (71.62) and "North-East" (73.43)). The word "south" appears as adverb, reflecting on graphic level that it is more often used as in "heading south" than referring to the "south of Birmingham". Also, the noun "city" (sig. of 154.53) is often related to the "city centre" (indicated through the very strong link between those two words), but also to make statements like "Birmingham is a city that" or "like other cities, Birmingham has". The references reveal the quality of this explanations, rather stressing its ordinariness as city instead of personalizing it or emphasizing its uniqueness. This indicates that the city itself is not standing prominently in the foreground in its crime novels (in contrast to Glasgow and in accordance with our qualitatively derived prior results). The proper name is mainly used as part of other proper names (e.g. "Birmingham Sentinel"), fulfilling the function of simply placing the plot, and there is very little city-specific information given on a statistical significant re-occurring level in the closer surroundings of it. Even the statements about Birmingham as a city tend to downplay its singularity.

On the German side, the cities names co-occur with words from a wider range of word classes. For both cities, we find less co-occurring proper names: five for Frankfurt, only one of them referring to a city-specific aspect (the long version of the name "Frankfurt on the Main" (sig. of 585.09)); four for Dortmund (again, only one city-specific, the name of its soccer club "Borussia" (with only seven instances and a sig. of 41.93)). For both cities, the rest of the proper names is composed of names of other cities (in Frankfurt the two nearby cities "Offenbach" (139.49) and "Darmstadt" (105.73), and "Berlin", "Hamburg"); for Dortmund only cities from the same metropolitan area (the Ruhrgebiet), "Düsseldorf" (41.95), "Werne" (41.78) and "Duisburg" (39.42)). It seems that Dortmund is closely connected within the metropolitan area it is a part of, but looking at the references shows that only Düsseldorf plays a role across different crime novel series, while the rest mainly feature in one certain series (being rated as author-specifc).

In the case of Frankfurt, the nouns that co-occur (seven) either denote city-specific aspects (Flughafen (airport) (96.83) and Eintracht (the local soccer club with a sig. of 192.36)) or very general instances (December, Jahren (years)). A look at the statistical concordances, ordered according not only to their position around the node but also to their significance, displays that the noun "Kripo" (short form for crime investigation unit) on the -1 position is more often used than the first city-specific instance, with a significance of 564.58 (while "police" for Glasgow on the +1 position is relatively ranked lower). This prominent position of the crime investigation unit (interpreted as impact of genre-related aspects) indicates that there are many "police-procedural" crime novels in Frankfurt (which is true), giving insight into the sub-genre composition of the corpus. As with the English cities, the word "Stadt" (city; sig. of 245.63) co-occurs frequently, and as the references reveal it serves similar purposes: either to denote the political administration (the "Stadt Frankfurt") or in combination with further explanations of "how this city is" (as in Glasgow, but without personalization), or "Frankfurt is a city that", but in contrast to Birmingham not with a frequent downplaying of uniqueness. Additionally, we find instances where other cities are compared to Frankfurt ("a city that, unlike/like Frankfurt"). This seems to point towards a more flexible use of this combination resp. to a variety of ways of representation. Frankfurt is represented as a city allowing for different semantizations and different ways of depicting it without posing contradictions (as the differing uses occur not only across a wide range of authors, but within the same texts).
Finally, taking a closer look at Dortmund, the frequently co-occurring nouns nearly all are related to genre-specific instances, referring to crime investigation-related institutions (again "Kripo" (sig. of 88.91); "Polizeipräsidium" (police headquarters; sig. of 35.15), "Landgericht" (district court; 37.25) and "Sonderstaatsanwaltschaft" (34.63)). This indicates that in this corpus the genre-specific structures seem to imprint themselves more than the city-specific ones, putting the city itself into the background (similar to the case of Birmingham but with a highly differing pattern). But we also have to consider the comparably low relative frequency rates (ppm) that demand an explanation. There might be another similarity between Dortmund and Birmingham, both showing low relative frequencies for their respective proper names. But as we take a closer look on the references of the occurrences of the names, we can see that the one series of crime novels that represents the biggest share of the corpus (with 21 novels belonging to this series) does not mention "Dortmund" at all, while for Birmingham the use of the proper name is quite equally distributed across all authors and series. A look inside one of this books of the series in question reveals a possible answer to the low frequencies: instead of using the proper name, the author consequently uses the nickname "Bierstadt" (Beer-city). Therefore, while it is possible to show that each city under investigation reveals a specific pattern of co-occurrences and differing uses and functions of its proper name, as our hypothesis has suggested, the search for the proper name alone seems not sufficient to get the overall picture of the literary representation of a city, demanding further analysis.

4.3 Analysis 2: Investigating Genre Aspects

When it comes to questions of genre-conventions vs. city-conventions, the investigation of the semantic preference of typically crime-related words is interesting. If the specific city has an impact on genre-aspects, the graphs should show clear differences. Close reading of exemplary novels of Glasgow and Birmingham indicated that violence plays a greater role in Glasgow crime fiction than in that of Birmingham, therefore we expect to find differing attributions towards and meanings of "violence", showing a higher vocabulary richness in Glasgow than in Birmingham, taking into account its semantic preference (for more details about this aspect see e.g. (Hoey, 2007)). We examine this hypothesis through making "violence" the node of a search for significant concordances, searching for the top-30 adjectives directly altering the noun within a range of -3 to +3 around the node.

As depicted in Figure 5, our initial hypothesis can be verified. While Glasgow (upper) has nine significantly co-occurring adjectives (six directly altering the noun "violence" on pos. -1), Birmingham (lower) only has five (four on pos. -1). Those that directly alter the noun show a slightly differing seman-

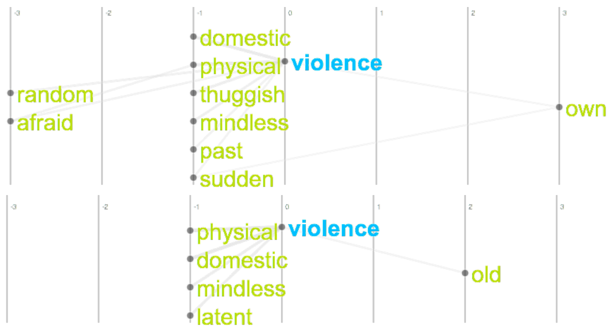


Figure 5: Significant adjective concordances of "violence", comparing Glasgow (upper) and Birmingham (lower) corpora

tic preference, with adjectives of "kind of violence" (domestic, physical) standing on top in both corpora. Next, we look at adjectives that bear a notion of "quality or intensity of violence": while Birmingham only discriminates between mindless and latent violence, the vocabulary of Glasgow is much richer (thuggish, mindless, sudden), one of them also bearing a notion of expectability (sudden). Additionally, a temporal adjective is used to refer to "past violence". If we look at the instances on the -3 position for Glasgow (a position that is not filled for Birmingham), we can add random to the list of "quality of violence", and find some instances of "being afraid of (physical) violence" (as the link between those words implies). This verifies our close reading interpretations.

The adjectives to the right of the node ("own" on position +3 in Glasgow, "old" on position +2 in Birmingham) pose a puzzle. Through a look at the references for this instances, we can see that in the case of Birmingham, old is referring to victims of violence (old people), while the picture for Glasgow is split between violence of its own type (which then could be added to the list of quality-adjectives) and violence that one experienced on his own. Through the interconnectedness of the adjectives settled on different positions for the case of Glasgow and a look at the resources of the instances, we conclude that the patterns seem to be more established on city level (showing instances from varying authors for all adjective-noun combinations) than they are in Birmingham, where there are no cross-connections and the authors differ more among each other (with "physical violence" being the only combination that occurs across different authors, while all other adjective-noun combinations only appear within the work of a single author).

5 Conclusion and Further Work

To conclude the exemplary analysis, CoocViewer helps not only to explore large corpora but also to verify or relativize impressions from classical qualitative literary research. It opens up new ways of exploring topics, themes and relationships within large sets of literary texts. Especially the combination and linkage of co-occurrences and significant concordances simplifies the analysis and allows a finer-grained and more focused analysis than KWIC concordances or simple frequency counts. The possibility to distinguish between these two viewpoints on the data accelerates and improves the interpretation of results. Additionally, the comparison between corpora is much facilitated through the immediate visibility of differing patterns. Further work can proceed along a few lines. We would like to enable investigations of the wide context of co-occurrences through access from the references back to the whole crime-novel document. Further, we would like to automatically compare corpora of the same language on the level of local co-occurrence and concordance graphs to aid generating hypotheses. This will make a change in the interface necessary to support a comparative view. Furthermore, we want to extend the view of the original text (see Figure 2) in our tool by centering the sentences according to the selected word or words, as done in KWIC views. When clicking on a single word, this would lead to the normal KWIC view, but selecting an edge we then want to center the sentences according to the two words connected by the edge, which might be useful especially for the concordances.

The tool and the pre-processing software is available as an open source project[12] and as a web demo.

Acknowledgments

This work has been funded by the Hessian research excellence program *Landes-Offensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz (LOEWE)* as part of the research center *Digital Humanities*.

[12] `https://sourceforge.net/p/coocviewer`

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada.

Sahra Allison, Ryan Heuser, Mathhew Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: an Experiment*. Stanford Literary Lab.

M. Barlow. 1999. Monoconc 1.5 and paraconc. *International journal of corpus linguistics*, 4(1):173–184.

Helmuth Berking. 2012. The distinctiveness of cities: outline of a research programme. *Urban Research & Practice*, 5(3):316–324.

Douglas Biber. 2011. Corpus linguistics and the study of literature. back to the future? *Scientific Study of Literature*, 1(1):15–23.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection - monolingual corpora of standard size. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

John F. Burrows. 1987. *Computation into Criticism. A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon, Oxford.

John F. Burrows. 1992. Computers and the study of literature. In Christopher S. Butler, editor, *Computers and Written Texts*, pages 167–204, Oxford. Blackwell.

John F. Burrows. 2007. All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1):27–47.

Tanya E. Clement. 2008. 'A thing not beginning and not ending': using digital tools to distand-read Gertrude Stein's The Making of Americans. *Literary and Linguistic Computing*, 23(3):361–381.

Hugh Craig. 1999. Jonsonian chronology and the styles of a tale of a tub. In Martin Butler, editor, *Re-Presenting Ben Jonson: Text, History, Performance*, pages 210–232, Houndmills. Macmillan.

Hugh Craig. 2004. Stylistic analysis and authorship studies. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell.

Jonathan Culpeper. 2002. Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. In Ulla Melander-Marttala, Carin Ostman, and Merja Kyt, editors, *Conversation in Life and Literature: Papers from the ASLA Symposium*, volume 15, pages 11–30, Uppsala. Association Suedoise de Linguistique Appliquee.

Jonathan Culpeper. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of shakespeare's romeo and juliet. *International Journal of Corpus Linguistics*, 14(1):29–59.

Chris Culy and Verena Lyding. 2011. Corpus clouds - facilitating text analysis by means of visualizations. In Zygmunt Vetulani, editor, *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 6562 of *Lecture Notes in Computer Science*, pages 351–360. Springer Berlin Heidelberg.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In *48th Annual Meeting of the Association for Computer Linguistics*, pages 138–147, Uppsala, Sweden.

Bettina Fischer-Starcke. 2009. Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistc analysis. *International Journal of Corpus Linguistics*, 14(4):492–523.

Bettina Fischer-Starcke. 2010. *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. Continuum, London.

Carsten Görg, Hannah Tipney, Karin Verspoor, Jr. Baumgartner, William A., K. Bretonnel Cohen, John Stasko, and Lawrence E. Hunter. 2010. Visualization and language processing for supporting analysis across the biomedical literature. In *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6279 of *Lecture Notes in Computer Science*, pages 420–429. Springer Berlin Heidelberg.

Michael Hoey. 2007. Lexical priming and literary creativity. In Michael Hoey, Michaela Mahlberg, Michael Stubbs, and Wolfgang Teubert, editors, *Text, Discourse and Corpora. Theory and Analysis*, pages 31–56, London. Continuum.

David L. Hoover. 2001. Statistical stylistics and authorship attribution: an emprirical investigation. *Literary and Linguistic Computing*, 16(4):421–444.

David L. Hoover. 2002. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, 17(2):157–180.

David L. Hoover. 2008. Quantitative analysis and literary studies. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literary Studies*, Oxford. Blackwell.

Masahiro Hori. 2004. *Investigating Dicken's Style: A Collocational Analysis*. Palgrave Macmillan, Basingstoke.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*, pages 105–116, Lorient, France.

Martina Löw. 2012. The intrinsic logic of cities: towards a new theory on urbanism. *Urban Research & Practice*, 5(3):303–315.

Martina Löw. forthcoming. The city as experential space: The production of shared meaning. *International Journal of Urban and Regional Research*.

H. P. Luhn. 1960. Key word-in-context index for technical literature (KWIC index). *American Documentation*, 11(4):288–295.

Michaela Mahlberg. 2007. Corpus stylistics: bridging the gap between linguistic and literary studies. In Michael Hoey, Michaela Mahlberg, Michael Stubbs, and Wolfgang Teubert, editors, *Text, Discourse and Corpora. Theory and Analysis*, pages 217–246, London. Continuum.

Michaela Mahlberg. 2012. *Corpus Stylistics and Dicken's Fiction*. Routledge advances in corpus linguistics. Routledge, London.

C. W. F. McKenna and Alexis Antonia. 2001. The statistical analysis of style: Reflections on fom, meaning, and ideology in the 'Nausicaa' episode of Ulysses. *Literary and Linguistic Computing*, 16(4):353–373.

Franco Moretti. 2000. Conjectures on world literature. *New Left Review*, 1(January/February):54–68.

Franco Moretti. 2007. *Graphs, Maps, Trees. Abstract Models for Literary History*. Verso, London / New York.

Franco Moretti. 2009. Style, Inc. reflections on seven thousand titels (British novels, 1740-1850). *Critical Inquiry*, 36(1):134–158.

Franco Moretti. 2011. *Network Theory, Plot Analysis*. Stanford Literary Lab.

Hans-Michael Müller, Eimear E. Kenny, and Paul W. Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *Plos Biology*, 2(11).

Rosanne G. Potter. 1988. Literary criticism and literary computing: The difficulties of a synthesis. *Computers and Humanities*, 22:91–97.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. 2006. Exploiting the Leipzig Corpora Collection. In *Proceesings of the IS-LTC 2006*, Ljubljana, Slovenia.

Thomas Rommel. 2008. Literary studies. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literary Studies*, Oxford. Blackwell.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Michael Stubbs. 2005. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14(1):5–24.

Christopher Tribble. 2010. What are concordances and how are they used? In Anne O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 167–183, Abingdon. Routledge.

Dominic Widdows, Scott Cederberg, and Beate Dorow. 2002. Visualisation Techniques for Analysing Meaning. In *Fifth International Conference on Text, Speech and Dialogue (TSD-02)*, pages 107–114. Springer.