DOI: 10.31703/glr.2020(V-I).17 Cite Us URL: http://dx.doi.org/10.31703/glr.2020(V-I).17

Citation: Khan, A., & Rasul, S. (2020). Extraction of Semantic Domains Through Corpus Tools. *Global Language Review, V*(I), 153-168. doi:10.31703/glr.2020(V-I).17

Azka Khan* | Sarwet Rasul†

p-ISSN: 2663-3299 | e-ISSN: 2663-3841 | L-ISSN: 2663-3299 | Vol. V, No. I (Winter 2020) | Pages: 153 – 168

Extraction of Semantic Domains through Corpus Tools

Abstract:

The increased interest in the techniques of corpus linguistics in the first decade of 21st century was based on the most important premises, which are valid even today – investigation of larger datasets in less time. This article compares the results of different corpus techniques employed for exploring the dominant semantic domains in a corpus. These corpus techniques include use of word clouds, frequency lists and KWIC of a text. This study uses fictional discourse by Kamila Shamsie – namely Broken Verses (2005) – to illustrate the corpus methodology. In addition to different corpus techniques, this study also compares the usability of different corpus software for this purpose such as, Antconc (3.2.4), Nvivo 11, and Sketch Engine. This article will prove to be a good beginning point for the researchers exploring a text in any field of corpus linguistics and digital humanities.

Key Words:

CADS, Digital Humanities, E-Humanities, KWIC, Lemma, Semantic Fields, Stemmed.

Introduction

The widespread use of computer technology in the last decade of the previous century drastically increased the number and scope of computer-aided researches made in the fields of corpus linguistics. Surprisingly, even after the availability of huge amounts of computer readable textual data and numerous computer-assisted automatic text analysers, computer-aided text analysis is still not a common approach in the sub fields of social sciences and humanities. This article endeavours to show the benefits and hurdles of using (semi-) automatic text analysis technologies for making qualitative studies in the field of digital humanities. This article does not suggest that the hindrances or limitations have been completely removed though; it proposes that there is a dire need to unlock the potential opportunities by encouraging the innovative researchers of digital humanities to explore, adapt and modify the newly developed approaches to the tons of digital texts available these days. This article also voices the concerns in extracting the dominant semantic domain from a fictional discourse with the help of corpus tools. This study also presents a systemized form of the selected features of three computer software for making qualitative researches easier.

For many years now, computer-aided text analysis is not limited to just counting words. Many new corpus software help the researchers explore the qualitative aspects of the data too.

Having a clear corpus methodology for extraction of semantic domains is important in a two-fold manner: for the language researchers it helps to understand the meaning of the text in less time; for the computational linguists it provides help to go beyond the simple counting of the most frequent words towards more complex understanding of human language by computer systems. Thus, a conceptual understanding of the context, bridges the gap between quantitative and qualitative research designs which can eventually lead to more sophisticated automatic extraction of "meaning" from of a discourse.

* PhD Scholar, Department of English, Fatima Jinnah Women University, Rawalpindi, Punjab, Pakistan.
Email: azkakhan80s@gmail.com
† Associate Professor, Department of English, Fatima Jinnah Women University, Rawalpindi, Punjab, Pakistan.

Development of computer automated systems has helped to overcome many challenges faced by the researchers of digital humanities but working in natural languages is still not free from ambiguity and complexity and extraction of semantic domains remains a challenge for the social scientists even now.

Aim and objectives

The current research has two main aims. First, it discusses the application and comparison of different corpus techniques to establish the dominant semantic domains in any discourse. A novel by Kamila Shamsie titled *Broken Verses* (2005) is used as an example to illustrate the findings but the methodology is applicable to any corpus in the field of digital humanities and social sciences. The corpus techniques used in this research include word clouds, frequency lists of both stemmed and synonymous words and KWIC. Secondly, the potential benefits of using different corpus software for extracting dominant semantic domains in a discourse are also pointed out, mainly by discussing three computer software, Antconc (3.2.4), Nvivo 11, and Sketch Engine. This research is guided by the following research questions.

1. How can we extract dominant semantic domains from a literary text by using corpus techniques?
2. Which features of the selected computer software help in this context?

Structure of the Current Research

This article is structured in three distinct parts. The first part reviews the related researches in digital humanities especially focusing on corpus assisted discourse studies (henceforth CADS) as an example. This section also explains the need for a replicable corpus methodology in extracting semantic domains from the selected text. In the second part the three methods for extracting the dominant semantic domains have been discussed. These three methods include usability of corpus techniques, namely, frequency lists, word clouds and KWIC for extraction of semantic domains. This part also discusses the limitations and reliability of these corpus software. The last part of this article consists of concluding remarks about the three methods employed for the extraction of semantic domain.

E-Humanities/Digital Humanities and Computer-Aided Researches in Social Sciences

Digital humanities (generally represented as DH) is an emerging field of study at the intersecting boundaries of digital technologies, mainly computers, and different sub-disciplines of humanities. In DH the development of scholarship involves collaboration in transdisciplinary researches and demands teaching and publication of computationally engaged researches (Terras, 2011). Production and employment of new computer applications and techniques, allows the DH researchers to experiment with new teaching techniques and adapted research approaches (Burdick, et. al. 2012). Thus, cultivation of a two-way collaborative relationship between the humanities and the digital, results in the development of a new scholarship. Corpus linguistics is one such sub-discipline of DH rapidly flourishing by the use of innovative research methodologies. On one side it involves participation of computational linguists for development of computer software and on the other it relies on the verification and validation of these software by corpus linguists. Historically, digital humanities have been associated with fields other than linguistics, such as humanistic computing, media studies, social computing but since the turn of the century, corpus linguistics has gained a prestigious position owing to the innovative researches made in it.

Methodological Scepticism and Semantic Ambiguity in Computer-aided Analysis

Using innovative modes give the researchers new insights but poses methodological problems too. Distribution of immense amount of informative data distribution on the World Wide Web, emails, blogs, memos, articles etc. demands extraction of useful information quickly and at a low cost. Text mining, topic modelling, computational content analysis (CCA) and Computer Assisted Qualitative Data Analysis

(CAQDAS) are some of the areas which focus on refinement of automated computational methods for dealing with enormous amount of knowledge in DH (Pollak, etal. 2011). The biggest hindrance in dealing with natural language texts is the problem of ambiguity of meaning and semantic uncertainty. Very few automated text analysis software can claim to extract semantically correct information from linguistic texts. Extracting linguistic information requires knowledge of lexemes and lemmas, a sound grip on specific syntax of the texts and understanding of the contextual context (Pollak, etal. 2011). Although syntactic parsing is used to solve the problem of lexical ambiguity, the problem is still not solved completely. The point is illustrated by discussing two examples given by (Wiedemann, 2013). Consider the following two sentences in this context.

1. I have put the baby in the pen.
2. He runs the company.

The syntactic processing (POS tagging) will help the computer system determine that the word *pen* belongs to the noun category of lexemes. Similarly the word *runs* is categorized as a verb. However, when the software tries to extract the semantic information of these two words, semantic ambiguity and uncertainty cause a problem. There can be at least three possible meanings of the word *pen:* a writing tool, a female swan, or an enclosure where babies can be lift. Similarly the word *run* has two meanings: an activity of controlling or a physical action. A reliable automated text analyser should be able to correctly interpret such problems of semantic ambiguity. So far the automated text analysers available are not reliable for such semantic ambiguities of natural languages. Thus using computational techniques for extraction of semantic domains in DH is not without problems and demands human intervention to avoid misleading results. Therefore, the studies made in this field are relatively small scaled. Secondly, the conclusions of such studies cannot be generalised to a broader scale. Thirdly, the experts of natural languages need more explanations of the step wise statistical methods adopted in the computer based studies, even more so if they want to replicate the methodological framework.

I have mainly drawn examples from the field of corpus linguistics and discourse studies in the next section to discuss some of the researches made in the interdisciplinary field of CADS (Corpus Assisted Discourse Studies) by employing computer software to review the status of researches available.

Current Trends in Discourse and Corpus Linguistics

Corpus linguistic techniques help to reveal and analyse the recurrent linguistic patterns in any discourse in a way that is not possible intuitively. In the last decade of 20 century, corpus stylistics established itself as a new field of interest (Sinclair 1991, Stubbs 1996).One early influence on the corpus stylistic analyses is Halliday (1971) who suggested that analysing the use of transitive and intransitive verbs in *The Inheritors by* Golding can lead to induce literary meanings from the text. Halliday demonstrated that the unique usage of a grammatical feature influences the meaning and message of the literary text. Tracing this link between the grammatical feature and the hidden message or, in other words, the link between *form and content* is almost imperceptible intuitively. Corpus techniques can help the researchers to analyse large sample of writing by a single author in a little time and therefore, provide empirical proofs for the analysis of *form/structure* which eventually helps in understanding the *content/theme*. Halliday (1971) concluded his research by suggesting that excessive use of intransitive verbs for describing a Neanderthal tribe helped the writer to highlight the passivity and lack of innovativeness. These traits made the survival of the tribe impossible in the course of evolution. However, Halliday's analysis has received strong censure by Hoover (1999) for problems of replicating the research methodology by future researchers. Hoover (1999) considers Halliday's methodology lacking explicit documentation as well as transparency of analysis to other analysts for their own research work. Burrows (1987) extracts literary meanings of discourse from linguistic data by discussing the relationship between idiolects used by the protagonist and their personality traits. Examples of corpus stylistic analyses include Burgess (1999), Hardy & Durian (2000) and Tribble (2000). All of them adopted Burrows' (1987) methodology to understand the relationship between the usage of lexical and grammatical

words in literary discourse and the meaning of the data. While analysing a discourse, the linguistic sample under investigation needs to be understood in relation to the accompanying context. This is one main reason that so far, discourse analysis has not been defined as a universal set of procedures which could be formalised into a computer package (Antaki et al., 2003) and poses new problems. Nevertheless, the use of corpus techniques for analysing discourse is termed *a methodological synergy* by Baker (2006). This methodological shift allowed the corpus linguists and discourse analysts to access a large scale data for generating more quantitative evidence than the small-scale data used previously. Corpus techniques allow not only for exploring the traditional texts like newspaper articles/editorial and speeches but also newer mediated texts for example face book comments and tweets. So far, fictional texts have not been explored much by them. The main reason of this neglect seems certain methodological problems faced by the language researchers. Firstly, a corpus tool cannot differentiate between the reported and reporting speech. Recently a software called CLiC has been introduced to analyse the local textual functions in fiction but its use is limited to searching only Dickens corpus and a few other 19 century reference corpora (Mahlberg etal. 2016). Nevertheless, the interface does not allow uploading a new text. Secondly, it cannot identify which pronoun is used for which fictional character. Thirdly, the figures of speech like metaphorical meaning, irony and pun on words, which are of great importance for meaning making in fictional discourse, cannot be identified by the corpus tool. The gap is still there and literary texts are used as a sample mostly in the field of corpus stylistics. The next section discusses the researches already available in the fields of CADS.

Need for a Systematic and Replicable Linguistic Analytical Framework for Extraction of Semantic Domains

Owing to the few researches made by using corpus techniques, there is an increased need to fill the gap by proposing the replicable and systematic methodologies, especially to resolve the issue of semantic ambiguity.

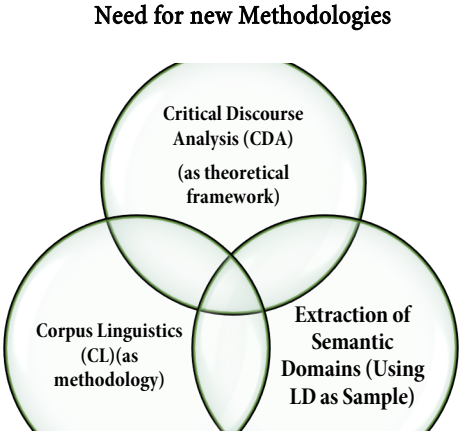

Fig 1: Corpus Methodologies for Extraction of Semantic Domains

Sally Hunt (2015) is one of those few researchers who analysed the process of representation of gender and agency in Harry Potter series by using corpus techniques. Hunt (2015) has focused on the words used for body parts of the social actors in this series. Since the field of CADS is in its incipient years, the choice of literary text selected for such a research is very important. Fischer-Starcke's work on *Pride and Prejudice* (2009) and Stubbs work on *Heart of Darkness* (2004 & 2005) are discussed as examples who give very important rationale for selecting these texts. Fischer-Starcke (2009) states that he has deliberately chosen a

novel which has been widely discussed and analysed for nearly last two hundred years by numerous critics. This makes the novel an especially attractive text for developing and verifying new corpus methodologies since it enables a comparison of findings by traditional methods of text analysis and findings by corpus based analysis. This helps the researcher to evaluate the effectiveness of the corpus techniques employed on the novel. The analyst can also focus on the linguistic/discursive processes used by the writer to construct meaning. Following the same rule Stubbs (2004 & 2005) used a century old novel *Heart of Darkness* for corpus stylistic analysis in which he tried to illustrate that the cultural and literary aspects of the novel can be shown with the help of frequency lists and distribution of words and recurrent phrases. This analysis also helped to identify important linguistic features which are usually missed by literary critics.

Extraction of Semantic Domains

Semantic domains as defined by Brinton (2001) are the groups of lexemes that share a common semantic property. Mostly these fields are defined by commonality of subject matter, such as landforms, colours, names of food items, or kinship relations. Computer-aided extraction of semantic domains from large amounts of texts can be useful in all the fields of Digital Humanities. Establishing credibility or high-precision in terms of methodology demands checking credibility of the tools and software available for corpus analysis. Extraction of semantic domains requires a three steps method:

(i) Syntactically categorizing the lexemes called POS tagging
(ii) Recognition of the lexemes from the same semantic fields
(iii) Clarifying semantic ambiguities (if any) to understand the relation between the selected lexemes and categorizing them semantically (Sematic tagging).

The reliability of some of the corpus techniques for extracting semantic domains available to the researchers of DH are discussed in the next section.

Employing Frequency Lists of Stemmed Words and Synonyms for Extracting Semantic Domains

An important principle, on which the foundation of corpus studies is laid, is the assumption that the most frequent lexical items are the most significant ones for establishing the dominant semantic fields and understanding the discourse structures (Sinclair 1991). Therefore, the frequency of lexical items is directly related to the structure and the content of the discourse. On the basis of this assumption the first corpus linguistic tool used in this research is to establish the dominant semantic fields are frequency lists. A novel by Shamsie titled *Broken Verses* is used as a sample in this research. The study corpus is abbreviated as *study corpus broken verse* (SCBV). While generating the frequency lists the functional words are not taken into the account believing that the main semantic load is carried by the content words. The software Nvivo 11 is used for generating frequency list because of its unique features discussed in the next section.

Unique Features of NVIVO 11

The unique features of NIVIVO 11 include the ease in uploading the corpus files. Nvivo 11 (Edhlund & McDougall, 2019) is a powerful software for qualitative data analysis which can run pdf. txt. rtf. and other files containing visuals and graphics. Unlike *Antconc* it does not require the study corpus (SC) to be changed into *TXT. format* prior to uploading it to the software. Another important feature in NVIVO 11 is that for generating the frequency lists, it automatically deletes the function words from the SC (Table 01 and 02). This way the researcher can focus only on semantically loaded words which are content words. This software provides two types of settings for generating the frequency lists.

1. Frequency lists may be generated by considering all the stemmed words as one entry e.g., *like, likes, liked, liking* etc. For the purpose of ease, in this research this list is termed as *Stemmed Freq. List* (see

table 01). The good thing in this setting is that it gathers all the lemmas of a lexeme as a single category. Thus the stemmed freq. list can be helpful in identifying the most frequent lexeme in a corpus (See table 01). The most frequent lexeme in SCBV is *mother.* Among the top twenty entries this is the only word which tells us something about the thematic content of the novel. The plot line in SCBV revolves around the most important character in the novel named Samina Akram. Her daughter Aasmani is the narrator of the novel and she uses the word *mother* very frequently for Samina Akram. Other than this word all the other words do not give any clue to the researcher for further exploration.

Table 1. Stemmed Freq. List of SCBV (top 20 entries)

Word	Length	Count	Weighted Percentage (%)	Similar Words
mothers	7	468	0.92	mother, mother', mothers, mothers'
ones	4	430	0.85	one, ones
just	4	383	0.76	Just
looked	6	363	0.72	look, looked, looking, looks
knowing	7	339	0.67	know, knowing, knowingly, knows
hands	5	275	0.54	hand, handed, handful, handing, hands
back	4	256	0.51	back, backed, backing, backs
years	5	240	0.47	year, years
even	4	237	0.47	even, evening, evenings
time	4	235	0.46	time, timed, times, times', timing
poet	4	234	0.46	poet, poet', poets
lovely	6	234	0.46	love, loved, lovely, loves, loving, loving'
days	4	220	0.43	day, days
think	5	210	0.41	think, think', thinking, thinks
way	3	202	0.40	way, ways
away	4	201	0.40	Away
want	4	194	0.38	want, wanted, wanting, wants
knew	4	191	0.38	Knew
never	5	185	0.37	Never
now	3	185	0.37	Now

2. The second setting used for generating frequency lists through NVIVO 11 involves categorizing all the synonymous words present in the text as one entry e.g., the most common word in SCBV is *look*. The software NVIVO has the ability to categorise all its synonyms under one head. Some of the words included in entry 01 Table 02 carry a very different semantic shade. To illustrate this point some words from the beginning of the list of synonyms are compared to the end of the list of synonyms. Words such as *appear, count, front, smell, sound, await* have many different shades of meanings (Table 02). The original entry *look* may be used as a synonym for these words but they are very different in meaning from one another. For example the word *appear* has a completely different meaning from the word *search* and *wait* has a completely different meaning from the word *smell*. This holds true for all the ten entries listed in table 02. Therefore, relying solely on the synonym Freq. list does not help a lot in the extraction of semantic fields. For the sake of brevity, top ten entries have been added to table 02. The words which have a very different meaning in the list of synonyms in front of each entry are put in the bold font.

Table 2. Synonym Freq. List SCBV (top 10 entries)

Word	Length	Count	Weighted Percentage (%)	Similar Words
looked	6	1158	1.33	appear, appearance, appeared, appearing, appears, aspect, attend, await, awaiting, bet, count, counted, counting, depended, depending, depends, expect, expectant, expectation, expectations, expected, expecting, express, expressed, expresses, expressing, expression, expressions, face, faced, faces, facing, feel, feeling, feelings, feelings', feels, front, fronts, look, looked, looking, looks, search, searched, searching, see, seeing, seem, seemed, seemingly, seems, sees, smell, smells, sound, sounded, sounding, sounds, spirited, tone, tones, wait, waited, waiting
mother	6	701	1.15	engender, father, fathers, fuss, generate, generated, generation, generation', generations, get, gets, getting, maternal, mother, mother', mothers, mothers'
Know	4	914	1.12	acknowledge, acknowledged, acknowledgement, acknowledgements, bang, banged, banging, bed, experience, experiment, experimenting, humps, intent, intention, intentions, intently, intents, jazz, know, knowing, knowingly, knowledge, knows, learn, learned, learning, learns, letter, lettering, letters, live, live', lived, lives, living, love, loved, lovely, loves, loving, loving', recognize, recognized, screw, wit, witness, witnessed, witnesses'
Going	5	1416	1.04	adam, become, becomes, becoming, belong, belonged, belongs, break, breaking, breaks, choke, choked, crack, cracked, cracks, departed, departure, die, died, dies, dying, endure, enduring, exit, exited, exiting, extended, extending, fail, failed, failing, failings, fit, fitted, fitting, flings, get, gets, getting, going, last, lasted, lead, leading, leads, leave, leaves, leaving, live, live', lived, lives, living, loss, move, moved, moves, moving, moving', offer, offered, offering, offerings, offers, operate, operating, operators, pass, passed, passing, plumpness, proceeded, proceedings, release, released, run, running, sound, sounded, sounding, sounds, spell, start, started, starting, starts, survive, survived, surviving, tour, touring, travel, traveller, travellers, travels, turn, turned, turning, turns, whirling, work, worked, working, workings, works
Just	4	703	1.03	bare, barely, exact, exacted, exacting, exactly, fair, fairly, good, goods, hard, hardly, just, justice, justify, mere, merely, precise, precisely, precision, right, righted, rightful, rightly, rights, scarcely, simply, upright

Think	5	889	0.99	believe, believe', believed, believing, conceive, consider, considered, considering, guess, guess', guessed, imagination, imaginations, imagine, imagined, imagining, intelligence, intelligent, intend, intended, mean, meaning, means, reason, reasonable, reasonably, reasons, recall, recalled, recalling, remember, remembered, remembering, remembers, retrieve, retrieved, suppose, supposed, supposing, think, think', thinking, thinks, thought, thoughtful, thoughts
One	3	467	0.88	one, ones, single, unity
Make	4	1104	0.86	attained, brand, build, building, buildings, cause, caused, causing, clear, cleared, clearly, clears, constitute, constitution, constitutional, construct, constructed, construction, cook, cooked, cooking, create, created, creates, creating, devised, draw, drawing, draws, earned, fashion, fashioned, fashions, fix, fixed, fixedly, fixing, form, formed, forming, forms, gain, gained, gains, get, gets, getting, give, gives, giving, hit, hitting, hold, holding, holdings, holds, make, makes, making, name, named, names, naming, piss, preparation, prepare, prepared, preparing, pretend, pretended, pretending, produce, produced, producer, producers, produces, producing, puddle, reach, reached, reaching, ready, realization, realize, realized, realizes, score, scored, scores, seduce, seduced, seduces, shit, shit', shuffled, stools, take, takes, taking, throw, throwing, throws, urine, work, worked, working, workings, works
Hand	4	496	0.73	custody, deal, dealing, fist, fistful, fists, give, gives, giving, hand, handed, handful, handing, hands, handwriting, men, pass, passed, passing, paws, reach, reached, reaching, script, scripts
Years	5	493	0.72	age, aged, ages, classes, day, days, year, years

For the sake of brevity, the complete frequency lists are not added here. Nevertheless the top twenty entries in the *stemmed freq.list* (table 01) and top ten entries in the *synonym freq.list* (table 02) make this evident that we need to apply some other corpus technique for the extraction of semantic fields. For this purpose the reliability of word cloud is discussed in the next section.

Employing word clouds as a beginning point to Extract Semantic Domains

A word cloud is commonly defined as a visualization of most prominent and frequent content words in a corpus. Word clouds are generated through frequency lists. The functional words are not added to word clouds as they reveal little about the semantic content of the corpus. They provide a low-cost and faster alternative than coding. Word clouds are generated on the basis of frequency by breaking the whole text into component words. The font point assigned to the words is directly proportional to the frequency of the word in the corpus. Word clouds have some benefits as well as some inadequacies as a corpus technique for revealing the semantic content of the corpus. It reveals only the essential information and provides an overall sense of the text. They have a visual appeal and are more

engaging than data in the stemmed tabloid form. The visual representation of word clouds generates interest but stimulates more questions than it answers. It can be a good entry point in a discussion about the data. The cons of word clouds in extracting semantic fields is that they can be misleading in interpretations. At times the size of equally frequent words is affected by the number of alphabets in a word or the size/shape of the glyphs. Randomly assigned coloured word clouds can also be misleading as some colours stand out more than others. Decorative fonts may have visual appeal but they sacrifice communication.

Word Cloud based on *stemmed freq.list*

Word Cloud based on *synonym freq.list*

Figure 2: Word clouds of SCBV based on *stemmed freq.list and synonym freq.list*

Two word clouds are generated for the SCBV, one is based on the *stemmed freq.list* while the other is based on *synonym freq.list.* Just like frequency lists the word clouds reveal little about the dominant thematic content of SCBV. In the next section reliability and efficiency of key words in context (KWIC) for extraction of semantic fields is discussed.

Employing KWIC (Key Words in Context) for Extracting Semantic Domains

List of keywords in context (KWIC) is different from simple frequency lists. Phillips (1985) suggests that keywords function to indicate the *'aboutness'* of the corpus. The keywords may not be the most frequent words of the study corpus, yet they are the most significant ones. Analysing the keyword list and categorizing the words according to their meaning reveal the dominant thematic content of the corpus. Scott in 2002 and more recently, Rayson (2008) and Culpeper (2009) have used this approach to reveal the meaning contained in various corpora.

Creating a Reference Corpus

Unlike frequency lists, word clouds, collocation lists and list of concordance lines generating KWIC requires a reference corpus (RC), in addition to the study corpus (SC). *Keyness* of any SC can be found out only by comparing it to another body of data. Some researchers (for example Sperberg-McQueen 1988) suggest that the keyword calculation of a sample text is somewhat effected by the RC chosen by the researcher. Others such as Baker (2006) and Stubbs (2005) suggest that by increasing the size of RC three times the size of SC, a keyword list free of any bias can be generated. There are two options available to all the researchers, either they can use the available large corpus as a reference corpus or they can build their own RC and feed it into software like Ant. Conc 3.5.8. Some software such as Sketch Engine and NIVIVO 11 have the in-built RC. In this research, English Web 2013 (enTenTen13) is available in the software Sketch Engine and is used to generate KWIC

Identification of the frequently occurring content-bearing lexemes in KWIC helped me derive the gist or *aboutness* or the dominant thematic content of SCBV. The KWIC are indeed the tip of the iceberg of meaning

but still provide reliable indications and manageable data for the detailed analysis of the main themes in the corpus. Instead of simple frequency lists, only the KWIC are focused for extraction of semantic fields in this section. The reason is that the word frequency lists are usually very long (reaching up to 2,041 items in SCBV) and the manual extraction of semantically relevant terms requires a lot of time. In order to make the length of target lists manageable the cut-off point is set 100 words. Table 03 contains the first 100 keywords of SCBV. The words which scored the highest in the *keyness* are the proper nouns. This is understandable because in *Broken Verses* most of the characters have Pakistani names that do not appear very frequently in the RC, thus these words qualify for a high score of *keyness.* The proper names do not tell us much about the semantic content of the corpus. Therefore, the names of the characters have been manually deleted from the list and after removing the names of the characters, top 100 keywords have been categorised and colour coded in Table 03.

Table 3. The Top 100 KWIC from SCBV

Top 100 KWIC of SCBV				
1	Single-word	Score	F	Ref F
2	Karachi	176.72	60	35,063
3	Laila	146.26	24	5,295
4	STD	132.12	49	40,431
5	Ramzan	125.58	18	1,796
6	Mama	121.63	87	98,938
7	Urdu	120.52	30	19,735
8	Qais	101.35	14	956
9	grazia	97.77	13	86
10	minion	95.05	34	38,260
11	Iblis	91.73	13	1,587
12	Macbeth	83.71	22	22,179
13	Hilal	80.78	12	2,782
14	Eid	79.39	20	20,340
15	Aadam	74.75	10	296
16	Archivist*	68.10	16	17,510
17	Fugue	64.09	11	6,775
18	Frass	58.70	8	806
19	Inqalab*	53.28	7	13
20	Ghazal	48.47	7	2,266
21	Shawl	47.89	18	41,579
22	Beloved	47.81	12	20,372
23	Lathi	44.69	6	583
24	Hikmet	44.69	6	583
25	Kabab	44.57	6	643
26	Nimue	44.38	6	743

27	Maulana	42.80	8	9,550
28	Zia	41.90	8	10,237
29	reshoot	41.86	6	2,156
30	schoolmaster	41.53	7	6,444
31	hoax	39.10	16	47,352
32	bougainvillea	38.92	6	4,038
33	Sadequain	38.28	5	48
34	Rafael	38.22	17	53,415
35	Hudood*	37.61	5	454
36	schoolfriend	37.35	5	613
37	mirage	36.35	8	15,277
38	Fata	35.94	5	1,528
39	Islamabad	35.27	12	35,692
40	mediaeval	34.33	6	7,620
41	crossword	34.17	10	27,635
42	Dad	33.83	73	344,417
43	ān (Quran)	33.56	5	3,249
44	dialled	33.39	5	3,385
45	impassioned	33.22	8	18,855
46	calligraphy	32.27	9	25,342
47	Amma	32.10	5	4,436
48	Morgana	31.92	5	4,586
49	stepmother	31.51	7	15,725
50	haiku	31.15	7	16,165
51	jalaibee	30.89	4	0
52	encrypt	30.88	24	110,022
53	captor	30.79	8	22,143
54	seekh	30.67	4	163
55	Sprezzatura,	30.64	4	187
56	falsa	30.56	4	248
57	maulana	30.54	4	259
58	resent	30.20	15	62,384
59	decrypt	29.83	7	17,889
60	Weep*	29.81	25	120,480
61	strangeness	29.50	6	12,581
62	iftar	29.02	4	1,462
63	absurdly	29.00	7	19,051

64	fizz	28.68	6	13,598
65	grandness	28.18	4	2,189
66	Ajar (open)	27.93	5	8,491
67	aur	27.78	4	2,541
68	punchline	27.56	5	8,908
69	Multan	27.52	4	2,778
70	kurta	27.38	4	2,914
71	ummah	27.32	4	2,970
72	mother	27.15	465	2,886,782
73	kameez	26.65	4	3,614
74	policewoman	26.62	4	3,641
75	Tyrant	26.07	4	4,202
76	unforgivable	26.00	5	10,811
77	newsreader	25.55	4	4,748
78	couplet	25.46	5	11,516
79	Gonzales	25.37	7	25,034
80	Bhutto	25.33	5	11,696
81	bookshelf	25.08	7	25,578
82	resentful	24.40	6	19,962
83	postmark	23.48	5	14,402
84	FUGUES	23.41	3	5
85	Nashaa	23.41	3	8
86	variedness	23.40	3	21
87	seventeen	23.39	13	72,642
88	Raqeeb	23.38	3	40
89	Frass	23.35	3	69
90	IMPRISONED*	23.33	3	85
91	sixteen	23.33	22	138,422
92	Mohtarma	23.27	3	143
93	chowkidar	23.26	3	157
94	calligraphed	23.11	3	299
95	Leucippus	23.07	3	340
96	Aashiq	23.07	3	341
97	unnaturalness	23.00	3	408
98	KDA	22.97	3	441
99	EXILE	22.93	3	484
100	gesture	22.87	43	297,602

The number of type and token of the first 100 KWIC in SCBV is calculated in the following way. The total number of top 100 keyword tokens are 1550. The total number of tokens in SCBV is 133,829 and total number of types is 10,288. The total number of *types and tokens* of the top 100 keywords from different semantic domains, their frequency, and percentage is given in the Table 04.

Table 4. Percentage of Token of top 100 KWIC in SCBV

	Semantic Fields/ Topic indicators	No. of Types in KWIC	No and % of Tokens in KWIC 1550	Definition and Comment	Most Frequent Examples from the Novel
2	Geographical locations	6	133 8.5%	To show the setting of the novel, there is frequent referring to Karachi and a studio STD	Karachi, Fata, Islamabad, Multan, KDA
3	Marriage and family life	7	674 43%	Familial ties and institute of marriage are a recurrent theme in SCBV.	Dad, mother, mama, beloved, stepmother, amma
4	Words from Regional Languages	18	150 9.6%	This category consists of words mainly from Urdu, and Punjabi.	Nashaa, Raqeeb, Mohtarma, Chowkidar, Aashiq, aur, Kurta, kameez, ghazal, Shawl, Lathi, Hikmet, kabab, Laila, Ramzan, Urdu, Qais
5	Political setup	11	63 4%	Many words included in this category needed the context to be reviewed and then they are put in this category.	Zia, Exile, captor, imprisoned, Tyrant Bhutto, archivist. Inqalab
6	An atmosphere of gloom and hopelessness	6	38 2.5%	The words in this category refer to negative feelings experienced by different characters but on the whole this group does not signify any one theme.	unforgiveable, resentful, unnaturalness, resent, absurdly
8	Miscellaneous	29	220 14%	Keywords not indicating any category	

The KWIC analysis helped to identify eight semantic fields out of which three categories, negative feelings, natural environment and miscellaneous did not help to signify a single theme. Figure 03 shows a graphic representation of the most dominant and the less dominant themes.

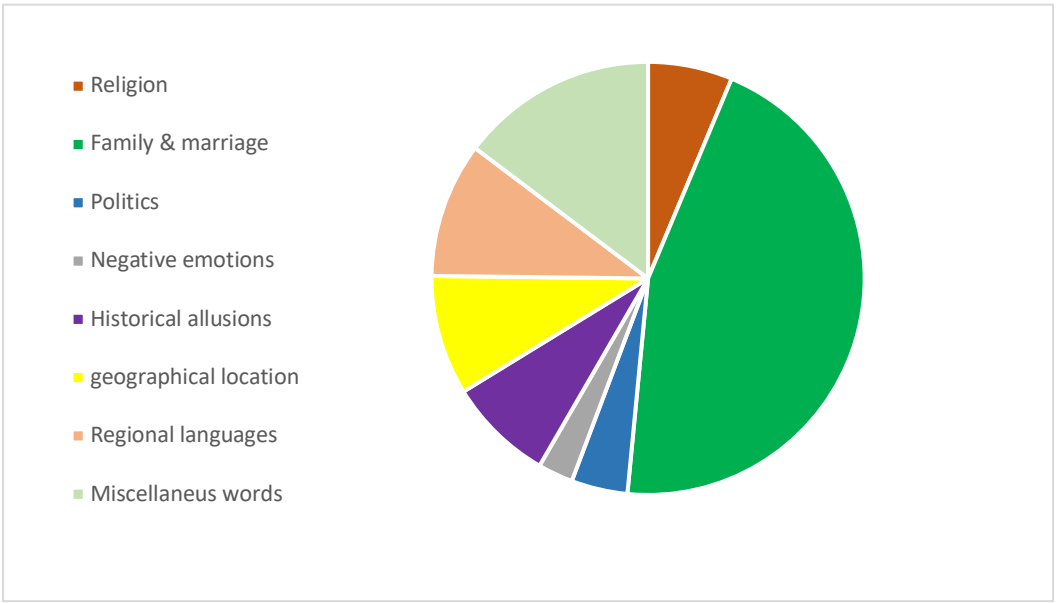

Fig 3: The Most Dominant and the Least Dominant Themes in SCBV

It needs to be made clear that some keywords are overlapping in terms of their thematic content for example a word which refers to an indigenous place can be put in either geographical locations or it can be taken as a historical reference. Similarly, the name of a regional language can be used for discussing a literary allusion. Therefore, figure 03 does not represent very clear boundaries; nevertheless, it does give an idea of the dominant themes in SCBV. It also shows the limitations of corpus techniques in terms of *aboutness* of the discourse. It is found that KWIC lists can give only a vague idea and blurred picture of the thematic content of the discourse and detailed collocation or concordance analysis is essential for understanding the detailed picture.

Some Methodological Concerns

Extracting semantic fields from SCBV through top 100 KWIC helped to gain the following methodological insights.

1. At the stage of categorizing KWIC into different semantic fields, I realised that I cannot rely only on KWIC for categorizing these words and the broader context of the words needs to be examined before categorising them into different semantic domains. Two examples has been given to illustrate this point. The word *heaven* occurs 8 times in SCBV. Superficially, it seems that this word belongs to the domain of religion but when the broader context is analysed, the findings were contrary to the initial expectations. This word is used *two times* for continuing the conversation in the phrase *for heaven's sake*. Similarly, the word *God* in SCBV is used as *thanks God, for God's sake, God forbid* etc. The researcher needs to note that these words are not actually referring to religion. On the other hand, some of the words such as *terror/ism, fundamentalist, radical and extremist* do not belong directly to

the semantic field of religion but when the broader context of occurrence is observed through concordance lines and paragraph retrieval, it is found that they are actually referring to religion.

2. Some words do not fit into any category. The category of words named *miscellaneous* in table 04 do not signify any one theme.
3. Some words with negative connotation (shown in grey colour in table 03, 04 and figure 03) in the corpus but they do not fit any one theme. It is still possible to conclude after concordance analysis of these words that the plot line is tragic or shows a gloomy atmosphere.
4. The code words used by Asmani (one of the main protagonists in SCBV) are recognised by the software as keywords because of their uniqueness but they do not reveal anything about the semantic content of the corpus so they are excluded from the list. These code words are *Ikrfb, fyfno, efac, Smaani, Anonkoh* are excluded from these lists.

Despite these methodological concerns, the use of KWIC for the extraction of semantic domains from a novel proved to be the most helpful when compared to all the other methods employed in this research.

Conclusion

This article demonstrated the use of three corpus techniques for the extraction of dominant semantic domains from a corpus. For this purpose, the fictional discourse produced by Shamsie titled *Broken Verses* has been used. The first two techniques, namely, frequency lists and word clouds can be used as the starting points to enter the data but they are not helpful in extracting the dominant semantic domains. The unique feature of the software NVIVO 11 is to produced frequency list based on synonyms also proved to be of little help due to the vast difference in the semantic shades of the words. The third method is consisted of manually categorizing the top 100 KWIC for extracting the semantic domains. This method is proved to be the most useful for the purpose of discourse analysis. The dominant semantic domains identified in SCBV through KWIC analysis are the same which are pointed out by literary critics after close reading of the texts.

References

Baker, P. (2006). *Glossary of corpus linguistics*. Edinburgh University Press.

Brinton, L. J. (2000). *The structure of modern English: A linguistic introduction*. John Benjamins Publishing.

Brinton, L. J. (Ed.). (2001). *Historical Linguistics 1999: Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9 13 August 1999* (Vol. 215). John Benjamins Publishing.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities*. Mit Press.

Edhlund, B., & McDougall, A. (2019). *NVivo 12 Essentials*. Lulu. com.

Hu, C. (2015). Using Wmatrix to Explore Discourse of Economic Growth. *English Language Teaching*, *8*(9), 146-156

Hunt, S. (2015). Representations of gender and agency in the Harry Potter series. In *Corpora and Discourse Studies* (pp. 266-284). Palgrave Macmillan, London.

Knowles, G., & Don, Z. M. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, *9*(1), 69-81.

Mahlberg, M., Stockwell, P., Joode, J. D., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, *11*(3), 433-463.

Pollak, Senja, Coesemans, R., Daelemans, W., & Lavrac, N. (2011). Detect ing contrast patterns in newspaper articles by combining discourse analysis and text mining. Pragmatics 21 (4): 647-

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519-549.

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment.

Rayson, P., Archer, D. E., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics conference: CL2007*.

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system.

Sharoff, S. (2004, May). Towards Basic Categories for Describing Properties of Texts in a Corpus. In *LREC*.

Stubbs, M. (2004). Conrad, concordance, collocation: heart of darkness or light at the end of the tunnel?' *The Third Sinclair Open Lecture*.

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

Terras, M. (2011). Quantifying digital humanities. UCL Centre for Digital Humanities.

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, 332-357.