# Unit 5 Case studies

## 5.1 Compiling Literary Corpora

Juliet Herring, University of Birmingham

This case study details the procedures and methods used in compiling a diachronic corpus of vampire literature, in this case 1800 to 2004. The study also describes the procedures used to create two additional reference corpora against which the main study corpus, henceforth the Vampire corpus, could be compared.

### Introduction

> Since electronic corpora became possible, linguists have been overburdened by truisms about the relation between a corpus and a language, arguments which are as irrelevant as they are undeniably correct. Everyone seems to accept that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within it and outside it that cause it to develop continuously. Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. (Sinclair 2004: online).

The aim of the project, of which this document only details the early data collection phase, was to investigate the linguistic features of Vampire Literature. It was thus necessary to collect data that was as representative as possible of the genre and organise it in such a way as to allow comparisons to be made that would consider the impact of the narrative structure and take account of such factors as time, author gender and narrator voice and, where possible, the gender and species of the narrator.

### Choosing the Source Medium

Before the age of the internet, computer based corpus compilation relied on texts being either manually entered by keyboard, or later, by scanning using Optical Character Recognition (OCR). This study, as its main source, used digitised texts sourced from the World Wide Web.

The use of digitised texts is obviously the quickest method of creating a large corpus but can impose some limitations:

1. Not all texts are available as electronic books (e-books)
2. Original text format/layout may have been lost
3. The e-book file format may not be compatible with corpus software
4. May contain transcriptions errors not found in the original text

Despite these limitations the use of digitised texts is still preferable to any other means currently available.

### The Vampire Corpus

The Vampire Corpus was the first of the three corpora to be created and its size and organization influenced the creation of the subsequent reference corpora. At the outset of the project there were no pre-conceived ideas about the genre in terms of what materials would be used and how they would be organised. It was during the early compilation phase that many of the features that would ultimately comprise the corpus were identified.

> When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested. (Renouf 1987: 2)

In order to fulfil these requirements, the development of the corpus was divided into seven key stages:

1. Identification of potential material.

2. Validation that it had been published in hard copy (i.e. as a book or in a magazine/newspaper).
3. Location of material.
4. Verification that the texts matched the 'vampire' criteria.
5. Converting texts, where necessary, to txt format.
6. Modifying the texts to remove author's/editor's notes, prefaces, copyright notices, advertising, etc and pre-processing to correct characters that would be mis-read by the corpus tool.
7. Organisation into sub-corpora.

To create the corpus it was first necessary to identify suitable material. As all the material required contained a key element, the vampire, it was a fairly simple matter to create a list of potential stories. For this, three different sources in addition to existing knowledge were used:

1. Internet Search Engines
2. Online Bookshop/Library Catalogues
3. *The Vampire Encyclopaedia*

Each source provided some information which was then cross-referenced. A database was then created using Microsoft Access to include various additional material.

The initial search used the used the key subject word *vampire* which produced a total of 6,240,000 hits, the term used to describe the number of sites matching the search criteria. The actual figure for hits for any word(s) varies each time the search is conducted.

This first pass list contained all occurrences found of the word which included such things as games, television programs, films, fetishist sites, bookshops, histories, Halloween costume stores, etc. To reduce the number of hits to a more manageable size and to remove some of the non-relevant hits, the list was filtered using selected context words:

- Story
- Tale
- Narrative
- Novel
- Text
- List
- Book
- Literature

Searching on each of these context words with the node *vampire* produced the following number of hits:

- Vampire + Story – 1,830,000
- Vampire + Tale – 499,000
- Vampire + Narrative – 57,600
- Vampire + Novel – 295,000
- Vampire + Text – 455,000
- Vampire + List – 1,710,000
- Vampire + Book – 1,520,000
- Vampire + Literature – 197,000

From these hits, sites were chosen for closer examination based upon their position in the list and on the summary details provided for the site.

The ease of access to the internet means that anyone can become a 'published' writer, establishing their own website. Selecting only texts that had been published in hard copy, acted as a nominal form of quality control. Although some authors now choose to write texts specifically for the e-book market and the inclusion of such texts would not necessarily affect the 'quality' of the material or adversely affect the results of the analysis, it was decided to

exclude them from the corpus. These texts may ultimately become a part of the mainstream genre but it could be argued that e-texts form a unique genre and that their inclusion could have actually skewed the results. According to Wikipedia, the first e-book publishing website went on-line in 1991. It can therefore be assumed that any publication date prior to 1991 refers to a hard copy and thus no further checks were necessary.

With a list of texts and authors, locating them became a comparatively simple matter. Entering the title of the text into the search engine, and then refining that search with additional parameters such as author name, publication date and file format together with the word *download*, identified sites where the material could be obtained.

The nineteenth century material, being out of copyright, is generally available for free on many web sites such as Project Gutenberg, the largest free online literary archive. Other specialist websites, such as Horrormasters.com, also provided material and although a nominal charge is usually made, the website owner kindly provided many texts free of charge for research purposes. All of those texts that were identified and could be obtained were included in the corpus.

The late twentieth and early twenty-first century material proved the most challenging. The growth of material in the genre made selection more difficult. Those that were available for free, and those that were inexpensive (less than $10.00) were generally selected for inclusion. However, given that this period is still expanding it was at some point necessary to stop adding texts and a cut-off date of 2004 was selected to allow the project to proceed to the analysis stage,. Additionally, given that there was already a heavy weighting within this period, it was decided that adding further texts would not necessarily provide a clearer understanding of the genre.

The story *The Skeleton Count* (Grey 1828), was found to be unavailable as an e-book and in order to make the nineteenth century as complete as possible it was felt that it should be included. The text was found only to have been reprinted in the anthology *The Vampire Omnibus* (1995) edited by Peter Haining. As it is a short story, it took little time to digitise the text using an OCR scanner. This method is not 100% accurate and some editing was later conducted where obvious errors were noted, such as '*a*' being read as '*e*'.

Vampire narratives can be found in many different genres including horror, gothic horror, dark fantasy, science fiction, fantasy and children's, to name but a few. However, not all fictional vampires follow the same model. For example, the story *The Flowering of the Strange Orchid* (Wells 1905) features a blood-drinking orchid; *Will* (O'Sullivan 1899) features a 'psychic vampire' who sucks energy without touching his victim, as well as a beetle that drains the life from the psychic vampire.

A decision was made to only include those texts that featured a vampire conforming to the standard mythical model of a blood-drinking 'human', often immortal or possessing an extended life span. It was thus necessary to read each text to ensure that it met these criteria. This process was aided by the use of Microsoft Reader, a text-to-speech facility, which effectively converted each text into an audio-book making it possible for simultaneous notes to be made for later reference.

During this process it was discovered that a number of the texts, while being listed as vampire stories, were in reality not and had been mislabelled. One example is the *The Scarlet Vampire* (Burke 1936). This verification process also aided in the location of errors such as duplicated paragraphs and extraneous content.

It was also decided at this stage to only include material from the genres of Horror, Gothic Horror and Dark Fantasy as it was felt that expanding into other genres would confuse the later analysis and make it difficult to faithfully compare the texts.

From the verified texts, not all were selected. When it came to the 1980 – 2009 sub-corpus, one author, Anne Rice, was felt to be over-represented. It was thus decided to limit an author's contribution to a maximum of four texts in any sub-corpus. This decision was not

based on any linguistic principle and nor was not possible to set the figure in terms of percentage of a sub-corpus as in some instances only one text was available in a sub-corpus. It was decided that bias that was created would be dealt with during the later analysis phase of the project.

Digitised texts can be stored in a number of different formats, identified by the file suffix, e.g. pdf, doc, lit, txt. Conversion to the required format, txt, was accomplished using commercially available software or by opening the file in its existing format and using the Save as function available in most software.

Digitised texts often vary from their hard copy equivalents and in addition, corpus tools often require an amount of pre-processing in order to be able to utilise the corpus effectively.

A test of the corpus with the corpus tool, WordSmith Tools 4.0 identified a number of problems that were corrected at this stage.

The first major problem that was identified by WordSmith was with the processing of the apostrophe.

> If your original text files were saved using Microsoft Word, you may find Concord can't find apostrophes or quotation marks in them! This is because Word can be set to produce "smart" symbols. The ordinary apostrophe or inverted comma in this case will be replaced by a curly one, curling left or right depending on its position on the left or right of a word. These smart symbols are not the same as straight apostrophes or double quote symbols. (Scott 2006)

This was corrected using WordSmith's own Text Convertor tool.

The texts exclude all prefaces, plot summaries and author/editor notes. Page numbers, where included in the downloaded text, were kept to be used later in referencing. Some converted texts retained additional information such as web source, author etc. This was particularly noticeable with pdf files. The use of the Find and Replace feature in Microsoft Notepad was generally used to locate and remove this information.

The story *Varney the Vampyre – The Feast of Blood* (Prest, 1847) raised two specific issues. The first concerned the chapter headings found in the files and as the second concerned the preface.

The story was originally published in three series as a *Penny Dreadful* and consisted of 237 chapters. Each of these chapters was downloaded as a separate file and found to be headed with the full title *Varney the Vampyre – The Feast of Blood* together with the chapter number and name.

At this stage it was assumed that the word *blood* would be significant within the genre and that the inclusion of 237 additional entries would distort the results. Later analysis confirmed this assumption and showed that *blood* was indeed a significant word. This was identified using the Keyword feature of WordSmith which compares the word frequencies found in a study corpus with a reference corpus using a mathematical formula. The program then identifies words that are statistically more frequent in the study corpus.

The second issue with the text was to do with the preface. In this preface the author addresses the reader directly stating that "the narrative is collected from seemingly the most authentic sources" (Prest, 1847) and goes on to state that "[n]othing has been omitted in the life of the unhappy Varney … and the fact of his death just as it is here related, made a great noise at the time through Europe, and is to be found in the public prints for the year 1713". This statement, that the story is factual, was considered to be an integral part of the narrative and was thus retained.

Some texts posed other problems. *Dracula's Guest* (Stoker), although first published in 1914 as part of an anthology of stories, is believed to be the original first chapter of *Dracula*

(Stoker, 1897) and was therefore included with the nineteenth century material and dated 1897. In addition it was considered to be a part of the journal of Jonathon Harker, although this is not indicated in the text and was therefore labelled as a first person male narrative (see below).

The vast majority of the texts were originally written in the English language although two translated texts were included. “Wake not the Dead” (Tieck, 1800 translated 1823), translated from German, and *The Dead Leman* (Gautier, 1836 translated 1903), translated from French, were both included in their English translation, the dates given and used for the corpus analysis, however, being the dates of the original texts and not the translation. In the case of *Wake not the Dead* this had no impact on the choice of sub-corpus. However, in the case of *The Dead Leman*, the translation date is substantially different from the original publication date. It was decided that although this could potentially contaminate the diachronic analysis the effect would be minimal and thus the original date and not the translation date was used for determining the appropriate sub-corpus.

Variant spellings such as *shew* and *show* were kept as were all the variant spellings of *vampire* (*vampir*, *vampyre* and *vampyr*).

The collected files were first divided into seven sub-corpora based on publication date:

- 1800 – 1829
- 1830 – 1859
- 1860 - 1889
- 1890 – 1919
- 1920 – 1949
- 1950 – 1979
- 1980 – 2009*

*In reality the last sub-corpus only covers the span of 1980 – 2004 in order to create a stable corpus for analysis.

By organising the data in this way, the corpus could later be analysed either as a whole or by historic period and similarities and differences between the periods could then be identified.

These sub-corpora were then further divided on the basis of author gender. The token count for these is given in the table below with the number of files shown in brackets. In later analysis comparisons at the level of author gender etc would be normalised by considering the results as a percentage of the relative sub-corpus.

| Historic Period | Token Count | Male Authored | Female Authored |
|---|---|---|---|
| 1800-1829 | 26,198 (3) | 18,875 (2) | 7,323 (1) |
| 1830-1859 | 625,473 (3) | 625,473 (3) | 0 |
| 1860-1889 | 72,690 (6) | 55,061 (4) | 17,629 (2) |
| 1890-1909 | 233,584 (14) | 233,584 (14) | 0 |
| 1920-1949 | 4,337 (1) | 4,337 (1) | 0 |
| 1950-1979 | 265,021 (4) | 132,901 (3) | 132,120 (1) |
| 1980-2009 | 3,636,232 (52) | 1, 131,428 (27) | 2,504,804 (25) |
| TOTAL | 4,863,535 (84) | 2,201,659 (55) | 2,661,876 (29) |

Table 1 – Vampire Corpus by Author Gender

The lack of female authored material became obvious at this stage and although a further search was conducted to locate material, none was identified.

Each file was also coded according to the following scheme:

- Narrative Type – Novel (N), Novella (A), Short Story (S)
- Voice – First Person (1), Third Person (3)
- Narrator Gender (only for first person) – Female (F), Male (M)
- Narrator Species – Human (H), Vampire (V), Other (only one example labelled Faerie)
- Narrative Name

In the case of Bram Stoker's *Dracula*, the file was subdivided to separate those sections narrated by male characters from those narrated by female characters. This resulted in two files labelled N1M-H-Drac.txt and N1F-H-Drac.txt respectively.

This coding system was to prove useful during the later stages of analysis.

**The Reference Corpora**

In addition to the main Vampire Corpus, described above, two additional reference corpora were created against which the Vampire Corpus could be compared. It was hoped that by creating two unique corpora for comparison the results that would be obtained later by keyword analysis would be more accurate.

The first reference corpus was a mixed genre corpus containing material from a variety of genres including Horror, Romance, Thriller and Westerns, to name but a few and referred to as the General Corpus. The second was a corpus containing texts from the same genres as the Vampire Corpus, namely Horror, Gothic Horror and Dark Fantasy and referred to as the Horror Corpus.

Each of these new corpora was divided into the same period based sub-corpora as the original Vampire Corpus. No further divisions were considered necessary at this stage as the effects of gender etc would be examined within the Vampire Corpus itself.

There were several key stages involved in creating the reference corpora.

1. Identification of Texts
2. Location of Texts
3. Division of Texts into Sub-Corpora
4. Using WordSmith to find the token count of each file
5. Matching the token count of each period as closely as possible

In order to begin the process of identifying materials for inclusion in these additional reference corpora a web search was conducted to find lists of the most popular fiction. These lists were based on sales figures for a particular historic period and thus gave no indication of publication date but did serve as a starting point. This method was also employed by the Cobuild project in defining the material to include in their corpus.

In addition to these lists, web sites were identified that supplied 'classic' texts. These latter types of site were most useful for locating short stories to fill-out the sub-corpora. Certain periods proved to be more difficult than others, most notably the 1800-1829 sub-corpus.

From these lists titles were located at various web sites. Where no list for a period could be found, well known texts and authors were targeted. For the nineteenth century Project Gutenberg was able to provide the majority of the material.

As with the Vampire Corpus, the major difficulty came with the more modern material which was again checked to ensure that it had been published in hard copy using the same check date. Cost was also a factor in choosing material as was author. In some cases texts were selected as they were by authors already represented in the Vampire Corpus. It was hoped that this would ensure that the keywords identified later by comparing the corpora, would be as the result of genre differences rather than author stylistic differences.

At this stage many more texts than would ultimately be needed were collected as their individual token counts could not be determined accurately based on file size alone.

These texts were then divided into the same 30 year periods as the Vampire Corpus using the first published date as the key.

The next phase was to try and match the individual sub-corpora sizes as closely as possible in terms of token count with the same period sub-corpus in the Vampire Corpus.

The token count for each file was determined using WordSmith. It was then a matter of mathematics to determine which combination of files would give the closest token count to the Vampire Corpus. It was at this stage that additional material was often needed and located using the methods already described. This sometimes meant choosing texts by less well known authors. Although it would have been possible to exactly match the sub-corpora sizes by using incomplete texts "it is not good practice to select only part of a complete artefact [and] it is an unsafe assumption that any part of a document or conversation is representative of the whole" (Sinclair, 2004 online).

The final result is that both of the reference corpora contain a similar token count to the original Vampire Corpus with an overall error of less than 1%, as can be seen in the corpora breakdown below. At the level of sub-corpus the difference is again less than 1% in all instances with the exception of the period 1920-1949. Here both of the reference corpora contain a higher number of tokens with the difference being 2.6% for the Horror Corpus and 2.1% for the General Corpus.

An alternative approach would be to consider each of the sub-corpora as a percentage of the appropriate corpus rather than consider it in terms of token count. These figures are given in brackets in the table below and again show how closely matched the corpora are. Using this method it would have been possible to make the reference corpora considerably larger and may have produced different keywords in the later analysis.

| Sub-Corpus | Vampire Corpus Token Count (% of Total) | Horror Corpus Token Count (% of Total) | General Corpus Token Count (% of Total) |
|---|---|---|---|
| 1800-1829 | 26,198 (0.54%) | 26,074 (0.54%) | 26,408 (0.54%) |
| 1830-1859 | 625,473 (12.86%) | 625,154 (12.86%) | 624,763 (12.85%) |
| 1860-1889 | 72,690 (1.49%) | 72,136 (1.48%) | 72,107 (1.48%) |
| 1890-1909 | 233,584 (4.80%) | 233,250 (4.80%) | 233,908 (4.81%) |
| 1920-1949 | 4,337 (0.09%) | 4,448 (0.09%) | 4,430 (0.09%) |
| 1950-1979 | 265,020 (5.45%) | 264,947 (5.45%) | 264,804 (5.44%) |
| 1980-2009 | 3,636,404 (74.77%) | 3,636,924 (74.79%) | 3,637,147 (74.78%) |
| TOTAL | 4,863,706 (100%) | 4,862,933 (100%) | 4,863,567 (100%) |

Table 1 – Token Count by Corpus/Sub-Corpus

**References**


Biber, D., S. Conrad, & Reppen, R. (1998) *Corpus linguistics: investigating language structure and use* Cambridge: Cambridge University Press

Bunson, M. (1993) *The Vampire Encyclopaedia* New York: Gramercy

Burke, N. (1936) *The Scarlet Vampire* London: Publisher

Gautier, T. (1836) *The Dead Leman* from http://www.horrormasters.com/Text/a0324.pdf

Grey, E. (1828) *The Skeleton Count* in Haining, P. (1995) *The Vampire Omnibus* London: Orion

Haining, P. (1995) *The Vampire Omnibus* London: Orion

Horrormasters. http://www.horrormasters.com/

McEnery, T. & Wilson, A. (1996) *Corpus Linguistics* Edinburgh: Edinburgh University Press

O'Sullivan, V. (1899) *Will* from http://www.horrormasters.com/Text/a0530.pdf

Prest, T.P. (1847) *Varney the Vampyre – The Feast of Blood* from http://varney.50megs.com/

Project Gutenberg http://www.gutenberg.org/wiki/Main Page

Renouf, A. (1987) 'Corpus Development', in Sinclair, J.M. (ed.) *Looking Up* London/Glasgow: Collins ELT

Scott, M. (2006) WordSmith Tools 4.0 http://www.lexically.net/wordsmith/version4/index.htm

Sinclair, J.M. (2004) "Developing Linguistic Corpora: a Guide to Good Practice – Corpus and Text – Basic Principles." Retrieved 30/10/2006, from http://ahds.ac.uk/guides/linguistic-corpora/chapter1.htm#section7

Stoker, A. (1897) *Dracula* from http://www.gutenberg.org/etext/345

Stoker, A. (1914) *Dracula's Guest* from http://www.gutenberg.org/etext/10150

Tieck, J.L. (1800) *Wake not the Dead* from http://www.horrormasters.com/Text/a0320.pdf

Wells, H.G. (1905) *The Flowering of the Strange Orchid* from http://www.horrormasters.com/Text/a2258.pdf

Wikipedia http://en.wikipedia.org/wiki/Main Page