In P. Blumenthal et al eds (2014). *Les émotions dans le discours. Emotions in Discourse*. Frankfurt/Main: Peter Lang. 237-53.

Michael Stubbs

Patterns of Emotive Lexis and Discourse Organization in Short Stories by James Joyce

Résumé

Le présent article se propose d'analyser les fonctions que des mots donnés peuvent assumer dans la structuration d'un texte à partir du recueil de nouvelles *Dubliners* de James Joyce. L'intrigue extérieure des 15 nouvelles étant généralement peu développée, c'est l'intérêt porté aux émotions des protagonistes ainsi qu'aux connotations évaluatives et symboliques de certains mots qui prime. Des réflexions générales sur le lexique du recueil dans son ensemble seront complétées par des analyses plus détaillées de deux nouvelles, *Eveline* et *A Little Cloud*. Dans ces nouvelles, on retrouve de nombreux mots désignant des émotions qui se situent tant à un niveau superordonné (p. ex. *feeling, mood*) qu'à un niveau spécifique (p. ex. *anger, remorse*). L'analyse du contenu textuel se basera sur des calculs de la fréquence de ces mots d'émotion, effectués à partir du texte brut ainsi que du texte sémantiquement annoté. L'analyse de l'organisation linéaire des nouvelles choisies se fera sur la base de calculs distributionnels, à partir desquels seront par ailleurs élaborées des représentations graphiques de la structure textuelle.

Abstract

This chapter studies the role which words play in structuring texts in the short story collection *Dubliners* by James Joyce. All fifteen stories in the collection have limited external action: what is of interest is the emotions of the central characters, and the evaluative and symbolic connotations of certain words. Generalizations about lexis in the whole collection are followed by more detailed examples from two individual stories, *Eveline* and *A Little Cloud*. The stories contain many words for emotions, both superordinate (e.g. *feeling, mood*) and specific (e.g. *anger, remorse*). Statistics on their frequency, based on raw text and on semantically tagged text, are used to study textual content. Statistics on their distribution are used to study the linear organization of individual stories, and are converted into graphic representations of textual structure.

1. Corpora and texts

This chapter discusses how the frequency and distribution of words for emotions and mental states contribute to narrative structure in two stories in the short-story collection *Dubliners* by James Joyce.

An empirical investigation of words in a given semantic field such as "emotions" can start from large corpora or from individual texts, but should ideally combine both corpus and text analysis. A corpus analysis typically studies recurrent collocations and lexico-grammatical patterns in a sample of many unrelated texts, produced by many different speakers. This is usually done by removing small fragments of language from their original source texts, and reordering them in an artificial way in a KWIC concordance. This is a valuable estrangement device which allows us to see new things. For example, we can compare the phraseology of a single text with typical phraseology in general language use. The last words of one of the stories in *Dubliners* are "… tears of remorse started to his eyes". If we search for the variable pattern "… tears of [EMOTION] VERB-ed PREP … eyes", we discover that it is not unique to this single story (these examples were collected from the world-wide-web via WebCorp: RDUES 1999-2012):

tears of agony	*rolled from*	*their eyes*
tears of delight	*welled in*	*his eyes*
tears of despair	*burst from*	*his eyes*
tears of desperation	*sprang into*	*her eyes*
tears of gratitude	*started from*	*his eyes*
tears of helplessness	*welled up in*	*his eyes*
tears of self-pity	*escaped from*	*her eyes*

However, artificially generated concordances and naturally occurring texts place fragments of language in fundamentally different kinds of context (co-text). If we place many small fragments of language in parallel but very local contexts (in a concordance), we can study their average use across a speech community. But such data provide no evidence about how such fragments contribute to the organization of whole texts. We need a different approach to study how words acquire symbolic and emotive meanings due to their frequency and distribution in specific texts, and how the uneven distribution of semantically related words (e.g. words for emotions) contributes to textual structure.

2. *Dubliners*: main themes

Dubliners (Joyce 1914) is a collection of fifteen short stories written in the early 1900s. In a letter of 1904, Joyce famously expressed his intention to portray the paralysis of Dublin and its citizens, and this theme has been discussed in detail by literary critics.

All the stories are superficially mundane accounts of city life, which contain no exciting plots or dramatic external events. Indeed, there is often little external

action at all, and the stories often come to no clear conclusion. For example, in *Eveline*, the main character plans to elope to South America with her boyfriend, considers the pros and cons during the whole story, but decides to stay in Dublin. In a word: nothing actually happens! The interest of the stories is in their social criticism of a city which is paralyzed by miserable social conditions, by British imperialism, and by Catholicism, and in their psychological criticism of characters whose obsession with their own emotions alienates them from others. They want to escape from their unadventurous lives, but are unable to act. They sometimes, at the end of the stories, show a moment of increased self-awareness.

3. *Dubliners*: narrative structure

Stories have beginnings, middles and ends. An initial indication of how word distribution contributes to the narrative structure of the stories in *Dubliners* is evident from a comparison of their opening and closing sentences. Most of the stories start rather traditionally, with opening sentences which refer, with little or no evaluative vocabulary, to the external world: times, places and characters. In contrast, the final sentences refer in an evaluative way to the internal feelings of the characters, and contain superordinate terms for emotions (e.g. *emotion*, *feeling*, *mood*, *sensation*, *sentiment*) and words for specific emotions (e.g. *anger*, *anguish*, *hatred*, *love*, *regret*, *remorse*, *shame*) or mental states (e.g. *approval*, *pain*, *recognition*). Here are clear examples from seven of the fifteen stories:

An Encounter:
[…] Every evening after school we met in his back garden. […]
[…] I was penitent; for in my heart I had always despised him a little.

Araby:
North Richmond Street […] was a quiet street. […]
[…] my eyes burned with anguish and anger.

Eveline:
She sat at the window. […]
[…] Her eyes gave him no sign of love or farewell or recognition.

The Boarding House:
Mrs Mooney was a butcher's daughter. […]
[…] Then she remembered what she had been waiting for.

A Little Cloud:
Eight years before he had seen his friend off at the North Wall [Dublin harbour].
[…]
[…] felt his cheeks suffused with shame […] tears of remorse started to his eyes.

Clay:
[…] The kitchen was spick and span. […]
[…] Joe was very much moved. […] his eyes filled up so much with tears. […]

A Painful Case:
Mr James Duffy lived in Chapelizod [an area of Dublin]. […]
[…] He could not feel her near him. […] He felt that he was alone.

I will show below that it is not a coincidence that the word *eyes* occurs in four of these examples. The final sentences in several of the stories contain further references to seeing (*gazing*, *glaring*, *looking*, *staring*, *watching*).

The patterns are tendencies, are not repeated mechanically, and some stories do contain words for emotions in their opening sentences, typically combined with more matter-of-fact descriptions of the external setting. For example, the first sentences of *Eveline* contain several references to time (*evening*) and place (*avenue*, *last house*, etc.), but also to Eveline's perceptions and mental state (*watching*, *odour*, *heard*, etc.):

> *She sat at the window watching the evening invade the avenue. Her head was leaned against the window curtains and in her nostrils was the odour of dusty cretonne. She was tired. Few people passed. The man out of the last house passed on his way home; she heard his footsteps clacking along the concrete pavement. […]*

In addition, words such as *eyes* and *head*, and also psychological verbs and verbs of perception (e.g. *see*, *know*, *looked*) are often frequent in fictional texts (Stubbs & Barth 2003, Stubbs 2005).

Nevertheless, the frequency and distribution of words from the semantic field of emotions contribute to the narrative structure of the stories in *Dubliners*. Key words occur at key positions.

4. Two stories: *Eveline* and *A Little Cloud*

The two stories *Eveline* (fewer than 2,000 words) and *A Little Cloud* (fewer than 5,000 words) share several themes. The action, such as it is, takes place within a few hours in Dublin in the early 1900s.

Eveline is the story of a young woman. She sits gazing out of the window of her dusty old house. She has planned to elope to Buenos Aires with a sailor named Frank, and imagines what life could be like far from Dublin. Her emotions alternate between a desire to escape her boring existence, her feeling that her home is not so bad after all, and a paralyzed state where she is unable to express any emotion whatsoever. She remembers incidents from her childhood: her domineering father, her mother's death. She goes to the harbour: Frank is waiting for her, but she cannot move or speak, and is frightened to leave Ireland.

A Little Cloud is the story of two friends who have not seen each other for several years. They meet in a bar and talk. Thomas Chandler has stayed in Dublin: he imagines what life could be like if he could read the reviews of the poetry which we know he will never write. Ignatius Gallaher has left Dublin; he does write and has become a "brilliant figure" in the London Press. Chandler's emotions alternate between superiority and envy. He is a clerk and, because of his better education, he feels superior to Gallaher, whom he finds "vulgar" in the way

that he boasts about his foreign travels. But he is also jealous of Gallaher's experience of foreign places. In a final scene, Chandler is back at home with his wife and child.

The minimal external action in the stories is based on banal cultural stereotypes: a young woman plans to elope, and a man has a row with his wife after an evening drinking in a bar. The reader must therefore be assumed to have the literary competence to recognize the convention that the stories are about something else: something significant about human experience (Culler 1975, 114). In each case, it is the emotions of one central character – and the implicit evaluation of these emotions – which are of interest: envy, frustration, insecurity, shame, vanity, a general dissatisfaction with life, but also an inability to do anything about it, and a final moment of (possible) insight. This literary competence also involves the ability to recognize repeated lexical patterns.

The rest of this chapter is a test of whether computational methods can give such observations an objective empirical basis, and whether they can discover anything new about a literary text which has been intensively studied for a hundred years (see also Stubbs 2005).

5. Raw word frequencies, relative word frequencies (keywords) and range

Since the vocabulary of a language is very large, almost all content words are very rare, and lexical patterns are often impossible to observe without techniques which can identify aspects of word frequency and range.

Word-frequency lists provide a rough confirmation that the external action in the stories is minimal and banal. The raw frequency of word-forms is the simplest statistic. Relative word frequency can also be informative: so-called "keywords" are words which occur in the text more frequently than would be expected, given their frequency in a large general reference corpus (Scott 1996-2012, Scott & Tribble 2006, Rayson 2012). Here, I have used the 100 million words of the BNC (British National Corpus).

The following list gives, for the whole of *Dubliners*, the content words which occur in both the top 50 word-forms (raw frequency) and the top 50 keywords:

> *said, man, little, old, good, asked, room, went, young, face, came, began, eyes, head, street, seemed, table*

The list is in descending frequency order. It ignores proper names/titles and grammatical words. Keywords were sorted by log-likelihood (for the advantages of this test for corpus comparison, see Rayson 2012).

In the list, we have (as expected) words which are generally high frequency in fiction (*said, asked*), plus a few words which are indicative, in a very general way, of the topics of the stories: references to characters and to indoor and outdoor settings (*young*, *old*, *room*, *street*, *table*). But note also the words *face, eyes* and *voice*. In the top 50 words (raw frequency) but not in the top 50 keywords are other words which signal indoor settings (*house*, *door*, *hall*) and several words for mental states (*think*, *felt*, *thought*, *knew*).

Statistics on the frequency of words must be combined with statistics on their range: that is, how many of the fifteen stories they occur in. This comparison is

necessary, since, for example, the word *snow* is relatively frequent (20) in *Dubliners*, but occurs in only one story, *The Dead*. These are the words (in descending frequency) which occur amongst the top 50 content words, and amongst the top 50 keywords, and in every individual story:

man, little, young, face, eyes, street, voice

These three filters, raw frequency, relative frequency and range, provide empirical confirmation of the significance of words which I had already started to identify, on intuitive grounds, in the closing sentences of the stories: *eyes* (freq 96), plus *face* (101) and *voice* (69). They are not words for emotions, but facial expression and tone of voice are ways of conveying emotions, and they signal, in some way, themes in the whole collection.

Keywords are, by definition, text-dependent. If the words *eyes*, *face* and *voice* are significant, then clearly this is true only of these specific stories. In addition, *Dubliners* is not, of course, a series of stories about people's eyes. Describing word frequency and range cannot explain anything, but it can draw our attention to objective textual facts which require interpretation.

6. Keywords and key semantic domains

It might seem that keywords are free of observer bias since they are generated automatically via a statistical test. However, keywords software typically picks out dozens of potential keywords, from which intuition must select those that are "interesting" (Moon 2007). In addition, such software calculates probabilities only for individual words, usually by comparing (as above) the relative frequency of word-forms in raw unlemmatized text. It does not group these word-forms under lemmas or semantic fields. Thus, semantically related forms (such as *remember*, *remembered*, *reminiscence*, *memory*) are counted separately, and low-frequency variants are missed entirely. It may seem intuitively obvious that *eyes*, *face* and *voice* are semantically related, but the software does not know this and cannot calculate the keyness of lexical sets. Also, since isolated words are removed from multi-word expressions (such as *make_light_of* or *in_light_of*), they may be misanalyzed (if the analyst looks at isolated words in frequency lists and assumes that the word *light* is a reference to "light").

These limitations are tackled in software designed by Rayson (2008), which groups words into key semantic domains. Each word-form in a text is tagged for its part-of-speech and then for its semantic field. The semantic tag set is organized in a thesaurus, which consists of 21 major semantic fields, subdivided into 232 more detailed categories. The software can recognize around 37,000 word-forms and 16,000 multi-word units, and assign the content words to these semantic categories. For example, the top-level field "Emotion" contains subdivisions including: "happy, sad", "calm, violent, angry" and "worry, concern, confidence".

Rayson (2008, 529) admits that the categories are "coarse-grained" and estimates the accuracy of the semantic tagging at (only) 91 per cent. Nevertheless, the software can test hypotheses about significant semantic fields in the stories. For example, if we compare the complete text of *Dubliners* with a reference corpus of written imaginative prose (here approximately 222,500 words from a

sub-corpus of the BNC), this confirms some points about textual content and structure, and modifies others.

A comparison confirms that words in the category "Emotion" occur more frequently than in the reference corpus, but at a relatively modest confidence level ($p < 0.05$, log-likelihood). A sample of word-forms in *Dubliners* in this category is in the Appendix, and shows that many items occur only once each, and are therefore missed in a "keywords" approach, but that their cumulative semantic effect is considerable.

The software also confirms that the following semantic fields are very significantly more frequent ($p < 0.0001$) than in the reference corpus. Note again the occurrence of *voice, face* and *eyes*.

"Speech: communicative" freq 1255 (top forms: *said, voice, say, told*)
"Anatomy and physiology" 1060 (top forms: *face, eyes, head, hand/s*)
"Religion" 288 (top forms: *God, soul, priest, chapel*)
"Music" 255 (top forms: *music, piano, tenor, sing*)
"Language" 191 (top forms: *word/s, read, expression, accent*)
"Light" 77 (top forms: *light, lamplight, shone*)
"Darkness" 50 (only two forms: *dark, darkness*)
"Mental actions and processes" 31 (all forms: *memory/ies, dreamed/t, intellectual, mental/ly*)

The category "Parts of buildings" 368 (top forms: *room, door window, hall*) confirms the indoor setting of many stories: this is something which is obvious to any reader of the stories. More interesting is the category "Frequency" 221 (top forms: *again, often, repeated, every_year, every_morning, night_after_night*). This provides evidence, which is probably less obvious to a reader's intuition, of one way in which the monotonous nature of the characters' lives is expressed.

7. Cyclic procedures

The more we know about a text, the more we can find out about it. Once we have noticed that *eyes* is a keyword, we might notice additional symbolic references to perception, which signal the frequently confused mental state of the characters. For example, Eveline is "confused" by the way Frank talks to her, and when it gets dark, she can no longer see clearly two letters which she is writing:

> *The evening deepened […]. The white of two letters in her lap grew indistinct.*

As Chandler enters the bar, he is also "confused":

> *He looked about him, but his sight was confused by the shining of many red and green wine-glasses. […] he felt that the people were observing him curiously.*

Examples of characters who are confused and unable to see clearly occur in other stories, and illustrate the importance of intertextual references within the collection:

> *[…] how I could tell her of my confused adoration. […] It was a dark rainy evening. […] I was thankful that I could see so little. (Araby)*
>
> *[…] a mist gathered on his glasses so that he had to take them off and polish them. […] (The Boarding House)*
>
> *[…] his eyes filled up so much with tears that he could not find what he was looking for. […] (Clay)*

Software can draw attention to the frequency of the word *eyes*, but it is unlikely that software could be programmed to recognize repeated symbols which are expressed with considerable lexical variation in different stories. This characteristic feature of extended (symbolic and metaphorical) language in literary texts may place a limit on automatic semantic annotation.

However, it illustrates important points about text and intertext. First, we interpret the stories differently whether we read each story independently, or as one of a series of stories which refer intertextually to each other. Second, in general English usage, the word *eye(s)* has different meanings (many non-literal) which are based on prototypical phraseology, and these meanings can be listed in dictionaries. As Hanks (2013) puts it, any given word has only a potential meaning: this is a state. But the meaning which is activated in a given text is an event. Third, the ways of expressing emotions are open-ended. The following additional examples from *Dubliners* describe characters' emotions. They contain no words for emotions as such, though (3) and (4) refer, in conventional phrases, to conventional ways in which people express their mental states.

> *(1) his little beady black eyes were examining me. (The Sisters)*
> *(2) I met the gaze of a pair of bottle-green eyes peering at me. (An Encounter)*
> *(3) he banged his fist on the table. (Counterparts)*
> *(4) Mr Holohan began to pace up and down the room. (A Mother)*

8. Lexical "burstiness" and narrative structure

In this chapter, I use the term "range" to refer to the number of different stories in the collection in which a word occurs, and "distribution" to refer to uneven periodicity of occurrence within individual stories. Frequency lists give a rough indication of textual content. However, words in texts occur in bursts and semantic clusters, and this uneven distribution of individual content words and words from particular semantic fields signals textual structure (e.g. Katz 1996, Bondi 2007).

Almost all content words are relatively rare, so there is only a low probability of finding a given content word in a given text. Exceptions are a few very high

frequency "general nouns" (Mahlberg 2005), such as *time*, *people*, *way*, *years*, and, in fiction texts, words such as *said* (Stubbs & Barth 2003).

However, once a given content word has occurred in a text, then it is likely to occur again, typically within the next few sentences (Alford 1971, 82, Manning & Schütze 1999, 547). The uneven lexical periodicity of this local "burstiness" contributes to textual structure. Again, we have to distinguish between frequency and range. Some words are frequent and recur across many or all of the stories (e.g. *eyes*, *night*), and some are frequent but only in one story (e.g. *snow* in *The Dead*, where it acquires symbolic meanings). So, we have an average burstiness (represented by "keywords") across all fifteen stories, which indicates (roughly) topics of the whole collection, and a local burstiness within individual stories, which signals both topic and narrative structure. (Katz 1996 distinguishes slightly differently between "average burstiness" and "topical burstiness").

9. A visual representation of narrative structure

A word list in descending frequency order gives a rough indication of textual content. However, we can also design software which identifies when individual words occur for the first time in a text, and when bursts of several new words occur in close proximity to each other. The software reads through a text word by word and marks each "new" word when it occurs. At the risk of belabouring the obvious, being a "new" word is not a property of language use in general, but of a specific text, and therefore of textual organization (Covington & McFall 2010, 95). For example, the last two sentences of *Eveline* are

> *She set her white face to him passive like a helpless animal. Her eyes gave him no sign of love or farewell or recognition.*

The words for emotions and mental states (*passive*, *helpless*, *recognition*) occur here for the first time. The words *love* and *face* have occurred previously, but the word *eyes* has not. We can get software to mark these first occurrences:

> *She >set her white face to him >passive like a >helpless >animal. Her >eyes gave him no >sign of love >or >farewell or >recognition.*

Similarly, in the last two sentences of *A Little Cloud*, words for emotions (*shame*, *remorse*) occur for the first time, along with words semantically related to *face* (*cheeks*, *tears*). The very last word, *eyes*, has occurred before.

> *Little Chandler felt his >cheeks >suffused with >shame and he stood back out of the >lamplight. He >listened while the paroxysm of the child's sobbing grew >less and less; and >tears of >remorse >started to his eyes.*

Youmans (1991) proposes what he calls the "vocabulary-management profile" of a text. As a text becomes longer, the number of word-tokens (running words) rises, by definition, at a constant rate, but the number of word-types (different words) rises more and more slowly, as words are repeated. Writers must repeat "old" words in order to make the text cohesive: this places limits on its lexical

diversity. But they must also choose "new" words in order to introduce new topics. These opposing pressures operate over whole texts, and cyclically over smaller sections, to produce uneven distributions of old and new vocabulary. Since "new" words occur in bursts, the type-token ratio can be calculated as a constantly changing ratio, across moving segments (spans) of text, in order to give a visual representation of textual structure (see graphs in Youmans 1991, Stubbs 2001, 136*sq*).

On average, the type-token ratio tends to fall in the course of a text, as more and more words are repeated. *Eveline* is around 1835 words in length (word-tokens), but has only around 600 different words (word-types). But, when new topics are introduced, new words are used, and so the ratio rises and there are peaks in the graph. Figure 1 shows a general downward trend, but peaks towards the end, at A where Eveline's mother's final madness is described, and at B when Eveline has arrived at the harbour.

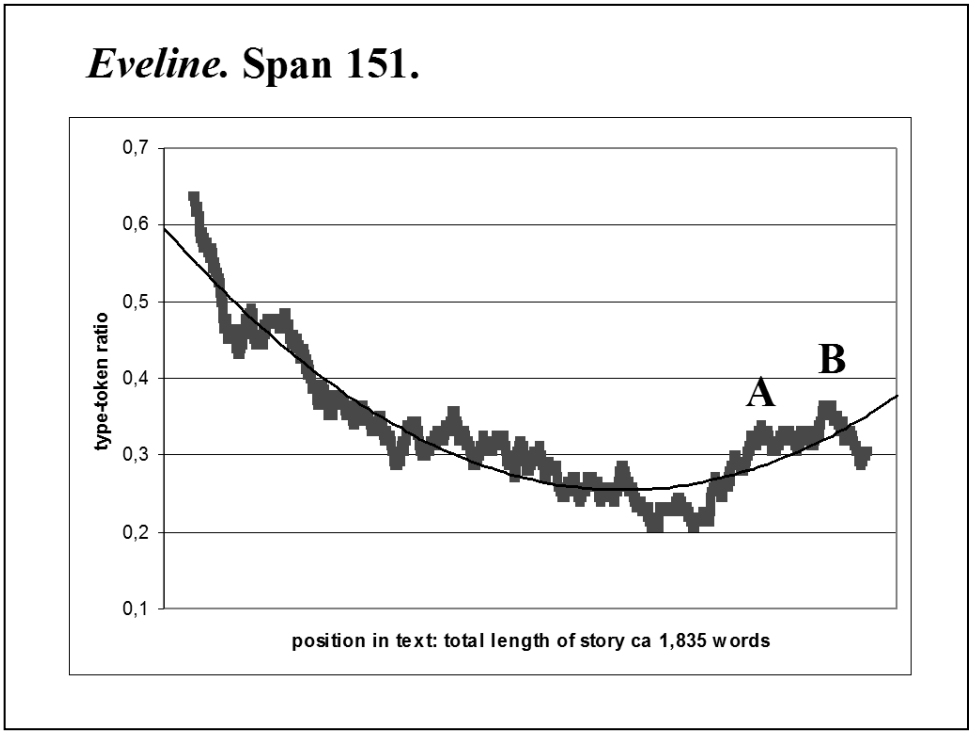

Figure 1: changing type-token ratios in Eveline, *span 151*

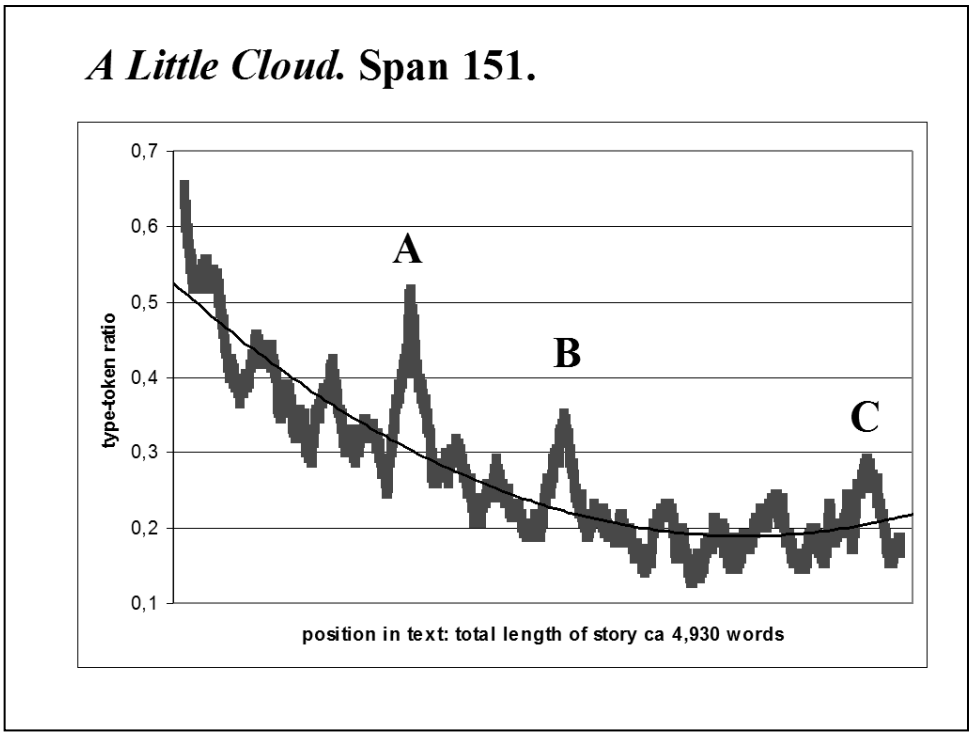

Figure 2: changing type-token ratios in A Little Cloud, *span 151*

Figure 2, for *A Little Cloud* (around 4,930 words), also shows a downward trend, but peaks at A when Chandler and Gallaher start to talk about their experiences since they last met, at B when Gallaher boasts about his adventurous life in cities abroad, and again at C near the end when Chandler is left alone with his baby.

10. A hypothesis: uneven word distribution and narrative structure

We can now formulate the following hypothesis. Towards the end of the stories are bursts of new vocabulary which signal a new topic. Much of this new vocabulary is from overlapping semantic fields: emotions and other mental states.

To test the hypothesis, we can compare the lexis in 200-word blocks at the beginnings and ends of *Eveline* and *A Little Cloud*. Since the semantic tagger embodies annotations which were designed independently of any personal interpretations which I might make of the stories, we can use the tagger to identify words in the category "Emotion" in Block 1 (at the beginning) and Block 2 (at the end). The comparisons provide corroboration of predicted tendencies in the lexical distributions.

The pattern in *Eveline* is clear. In Block 1, the tagger finds only one "Emotion" word (*happy*). In Block 2, it finds several "Emotion" words (*anguish*, *cry*, *distress*, *mournful*). In Block 2, it also finds several "Psychological" words, which could arguably also be categorized under "Emotion" (*fervent*, *felt*, *passive*, *recognition*). Most of these words occur for the first time in Block 2. However, a manual check identifies other words in Block 2 which also denote Eveline's emotions (*frenzy*, *heart*, *helpless*, *love*, *nausea*, and possibly *called*, *prayed*, *prayer*, *shouted*). The tagger codes these words under different semantic

categories. Coding individual words and multi-word units also misses longer expressions which convey Eveline's emotions, for example:

all the >seas of the >world tumbled about her heart
he would >drown her
she >gripped with both >hands at the >iron >railing

In this comparison, the tagger scores high on precision (it finds relevant words), but less well on recall (its findings have to be manually enhanced). However, it would be unfair to expect a tagger designed for general purposes to be able to handle entirely comprehensively a literary text from the early 1900s, and to identify such metaphorical expressions (which clearly cannot be listed in a thesaurus).

The pattern in *A Little Cloud* is also clear. In Block 1, the tagger finds only two "Emotion" words (*fearless, smiled*). In Block 2, it finds several, most of which occur for the first time in Block 2 (two each: *cried, cry, love, sobbing*; and one each: *fright, frightened, hatred, remorse*). Again, a manual check identifies further words in Block 2 which also denote emotion felt by Chandler and his child (*paroxysm (of sobbing), shame, panting, stammer*). And again, several relevant individual words are not identified because they are part of longer expressions which describe conventional ways of conveying the wife's emotions.

the door was >burst open
she >flung her >parcels on the >floor
>snatched the child from him
>giving no >heed to him
>clasping the child >tightly

Semantic categories are not sharply defined, and although semantic tagging in Wmatrix depends on a prior grammatical tagging, it cannot be entirely sensitive to the co-text. In *Eveline*, for example, *love* is categorized not as "emotion", but as "social action or state" (along with *helpless, prayed, prayer*). Similarly, other words which could be said to describe "emotions" are categorized as "psychological actions or states" (*felt, fervent, passive*), as "body" (*heart, nausea*) and as "general actions" (*frenzy*). Wmatrix does however distinguish between the use of the word-form *love* in *Eveline* (as "social action or state") and in *A Little Cloud*, where it is an address term (*Was 'ou frightened, love? There now, love! There now!*).

The general theoretical issue in my repeated point about the word *eyes* is as follows. Joyce's use of the word is by no means ungrammatical or "odd". In everyday English (as sampled in the BNC), eyes are *beady, dark* and *large,* and the word collocates with *face* and *tears*. But in *Dubliners*, by its frequency, range (across all the stories) and distribution (several times at the very end of individual stories), the word is given textual meanings in addition to its shared conventional meanings in the speech community (Hanks 2013).

Words can acquire symbolic meanings due to their frequency and distribution in individual texts. These textual meanings cannot be captured in a general lexicon (of the type which provides the basis of a semantic tagger). As Sinclair (2004, 21)

puts it, meanings are entirely provisional and are created by "ad hoc interpretation" in different texts. Hence, meanings are not something that can be recorded comprehensively in reference books and "no finite lexicon can include them all".

11. Binary contrasts and evaluative vocabulary

A systematic comparison of all fifteen stories, using the methods I have illustrated, will have to wait for a longer study. So will a discussion of other aspects of narrative structure and evaluative vocabulary which I can mention here only briefly.

Readers with basic literary competence will recognize the importance of binary contrasts (Culler 1975, 126) in many of the narratives in *Dubliners*. The point of *Eveline* rests on the contrasts between Eveline's daydreams and reality, between Eveline who lives on the edge of town and who is frightened to leave Dublin, and Frank who lives "in a house on the main road" and who has gone to Buenos Aires. The point of *A Little Cloud* rests on the contrasts between Chandler's Dublin and Gallaher's foreign cities, between Chandler's daydreams and reality, and between Chandler (who wants to be a poet but never will be) and Gallaher (who has become a journalist).

A series of contrasting adjectives, which all have evaluative connotations, signal this contrast. Dublin is *poor* and *dull*. Gallaher's places are *rich* and *lively*. Chandler is *little*, *refined*, *superior*, *childish*, *modest*, *sober*. Gallaher is *large*, *shabby*, *inferior*, *brilliant*, *wild*, *vulgar*, *gaudy*. A major finding of corpus analysis is the pervasiveness of evaluative language, and the extent to which the evaluative meaning of individual words depends on their typical collocates. For example, Chandler is envious of Gallaher but disillusioned by the way he has become "vulgar" and "gaudy".

> *There was something vulgar in his friend which he had not observed before. But […] the old personal charm was still there under this new gaudy manner.*

Corpus-based dictionaries (e.g. Cobuild 2009, LDOCE 2009) note the disapproving connotation of the word *gaudy*. And concordance data (collected via WebCorp) confirm that *gaudy* frequently collocates with *vulgar* (e.g. *brash, vulgar and gaudy*; *their vulgar fashion and gaudy antiques*; *loud, gaudy and arguably vulgar*; *such gaudy celebrations of vulgar wealth*). Such examples make the point again that the lexis which signals emotions and mental states is not restricted to labels for emotions as such.

The binary structure of the narrative is signalled explicitly in the text:

> *[Chandler] felt acutely the contrast between his own life and his friend's […].*

And – as one might perhaps expect, given my earlier discussion – one of the contrasts concerns the eyes of Chandler's wife and of Gallaher's women:

> *[Chandler's wife's] eyes irritated him […] there was no passion in them […] He thought of what Gallaher had said […] Those dark Oriental eyes, he thought, how full they are of passion. […]*

12. Summary and conclusion

This chapter has discussed a formal feature of language in use which can be tracked by software across texts and text collections: the repetition of individual word-forms and words from defined semantic fields.

There is no fixed set of words in English which label emotions, since the boundaries of semantic fields are fuzzy, and since words acquire evaluative meanings from their recurrent collocations in general language use and in specific texts. Some words (e.g. *gaudy*) have evaluative meanings which are conventional across a speech community, and which should therefore be listed in dictionaries. Others are encoded in idiomatic phrases in the language (e.g. *paced up and down the room*; *banged his fist on the table*): they are probably not good candidates for dictionary entries. Symbolic and emotive meanings which develop within individual texts (e.g. *eyes*) should clearly not appear in dictionaries.

This chapter argues that computer-assisted techniques can help to identify textual and intertextual features across a set of related literary texts and across these texts and large corpora, and is at some mid-point between (a) linguistic text analysis and (b) quantitative literary stylistics.

In terms of (a), it shows one way in which lexis contributes to textual organization. This is an area where there are many interesting case studies, but not yet a comprehensive functional theory of lexis.

In terms of (b), analyzing the frequency and distribution of words has drawn my attention to stylistic features of *Dubliners* which I had not previously noticed. Whether I have added to anyone else's interpretation of the book I cannot say.

Acknowledgements

I am very grateful to Paul Rayson, Gilbert Youmans and Peter Dingley for the use of their lexical analysis software, to Sabine Erschens for help with coding data, and to Gabi Keck, Amanda Murphy and Markus Müller for comments on previous drafts. I have used the version of *Dubliners* available from Project Gutenberg.

References

Alford, M. T. H. (1971). "Computer Assistance in Language Learning", in: Roy A. Wisbey (ed.): *The Computer in Literary and Linguistic Research*. Cambridge: Cambridge University Press, 77-86.

Bondi, Marina. (2007). "Historical Research Articles in English and Italian", in: Marcella B. Papi *et al*. (eds.): *Lexical Complexity*. Pisa: Pisa University Press, 65-83.

Cobuild (2009). *Collins Cobuild Advanced Dictionary*. Glasgow: Harper Collins.

Covington, Michael A. & McFall, Joe D. (2010). "Cutting the Gordian knot: the Moving-Average Type-Token Ratio", *Journal of Quantitative Linguistics* 17(2), 94-100.
Culler, Jonathan D. (1975). *Structuralist Poetics*. London: Routledge.
Hanks, Patrick (2013). *Lexical Analysis*. Cambridge, MA: MIT Press.
Joyce, James (1914). *Dubliners*. London: Grant Richards.
Katz, Slava M. (1996). "Distribution of Content Words and Phrases in Text and Language Modelling", *Natural Language Engineering* 2(1), 15-59.
LDOCE (^{5}2009). *Longman Dictionary of Contemporary English*. Harlow: Pearson.
Mahlberg, Michaela (2005). *English General Nouns*. Amsterdam: Benjamins.
Manning, Christopher D. & Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
Moon, Rosamund (2007). "Words, Frequencies and Texts (particularly Conrad)", *Journal of Literary Semantics* 36, 1-33.
Rayson, Paul (2008). "From Key Words to Key Semantic Domains", *International Journal of Corpus Linguistics* 13(4), 519-549.
Rayson, Paul (2012). "Corpus Analysis of Key Words", in: Carol A. Chapelle (ed.): *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.
RDUES (1999-2012). *WebCorp*. [Research and Development Unit for English Studies]. Birmingham City University, URL: <<http://www.webcorp.org.uk/live/>> [02.01.2013].
Scott, Mike (1996-2012). *WordSmith: Software Tools for Windows* [software]. Oxford: Oxford University Press.
Scott, Mike & Tribble, Christopher (2006). *Textual Patterns*. Amsterdam: Benjamins.
Sinclair, John (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.
Sinclair, John (2004). "Meaning in the Framework of Corpus Linguistics", *Lexicographica* 20, 20-32.
Stubbs, Michael (2001). *Words and Phrases*. Oxford: Blackwell.
Stubbs, Michael (2005). "Conrad in the Computer: Examples of 'Quantitative Stylistic Methods'", *Language and Literature* 14(1), 5-24.
Stubbs, Michael (in prep). "Quantitative Methods in Literary Linguistics", in: Peter Stockwell & Sara Whiteley (eds.): *The Handbook of Stylistics*. Cambridge: Cambridge University Press.
Stubbs, Michael & Barth, Isabel (2003). "Using Recurrent Phrases as Text-Type Discriminators", *Functions of Language* 10(1), 65-108.
Youmans, Gilbert (1991). "A New Tool for Discourse Analysis: the Vocabulary-Management Profile", *Language* 67(4), 763-89.

APPENDIX

The following is an illustrative sample of words (all words beginning with *a* to *c*) identified by Wmatrix (Rayson 2008) in the category "Emotions" in the complete text of *Dubliners*. A large number of the word-forms occur only once or twice, and would be missed in a search for "keywords". See, for example: *anger* 7,

angered 1, *angrily* 2, *angry* 6 and *blush* 4, *blushed* 2, *blushes* 1, *blushing* 3. The precision attained by the tagger is high: all the word-forms seem intuitively relevant to the expression of emotions. The recall is unknown: as noted in the chapter, other relevant words are captured under other headings such as "Psychological states, etc.".

abhorred 1, abuse 1, abused 1, adoration 1, affection 3, affectionate 2, affections 1, afraid 10, agitated 2, agitation 4, agreeable 1, alarm 2, alarmed 4, alas 1, amuse 2, amused 3, anger 7, angered 1, angrily 2, angry 6, anguish 2, annoy 1, annoyance 1, annoyed 5, anxiety 1, anxious 3, anxiously 1, applauded 3, applauding 1, applause 6, appreciated 2, appreciation 1, appreciative 1, apprehensively 1, approvingly 1, at_ease 5, attack 1, attacked 3, battered 1, bliss 1, blush 4, blushed 2, blushes 1, blushing 3, boldly 9, bother 2, brave 2, braved 1, bravely 3, bravery 1, bristled 1, brood 1, brooded 1, brutal 2, bullied 2, calm 4, calmed 1, calmer 1, calmly 3, care 1, cares 1, caressing 1, celebrated 1, cheerful 2, cheerfully 4, cheerfulness 1, cheerless 1, cherish 1, comedy 1, comic 3, comical 2, complacent 1, concern 2, confidence 2, content 2, contented 1, contentedly 1, courage 4, coward 2, cried 20, cross 1, cry 10, crying 4