A phraseological approach to film dialogue: Film stylistics revisited

MARIA FREDDI

Abstract

The paper takes film dialogue as a test case for a corpus-driven investigation of phraseology. The analysis is mainly based on three strands of research, both linguistic and translational. These are: corpus work comparing contemporary film and television dialogue with natural conversation, research on translational routines in audiovisual translation (dubbese) and the phraseological approach to language. However, in order to focus on the formulaic features of English original film dialogue, the translational perspective is backgrounded. The study is corpus-driven in that formulae of filmic speech are extracted from the corpus on the basis of their frequency. Furthermore, the sequences thus found are compared to general reference corpora of British and American English in order to explore their distribution and functions. The results are shown to be relevant to a stylistic appraisal of scripted film dialogue as well as to an understanding of some methodological issues associated with corpus-driven studies of phraseology in general.

Keywords: film dialogue; formulaic expressions; corpus-driven phraseology; film stylistics.

1. Background

Recent findings on the similarities between contemporary film and television dialogue and real-life unscripted conversations, together with research on translational routines in audiovisual translation, serve as a background to the present analysis of formulaic language in film dialogue. In particular, through systematic comparisons between the typical features of spontaneous conversation and television dialogue, Quaglio (2008, 2009a, 2009b) has shown how

TV dialogue tends to capture and reproduce the linguistic characteristics of authentic face-to-face conversations. Among the features he analysed are: first- and second-person pronouns, discourse markers (including *okay*, *well*, *you know*), intensifying adverbs (like *so* in *This is so not like him*, or *totally* in *I totally like it*), hedges (*kind of*, *sort of*), emphatics (e.g., *just*), slang terms, vocatives and familiarisers (e.g. *guys*), etc. These were studied in the very popular sit-com *Friends* and in natural conversation, and, although the two kinds of talk present different distributional profiles, they can be explained in terms of functional and situational differences. For example, features of emotional language are found to be consistently more frequent in *Friends* than in the conversation corpus, due to the dramatic nature of the dialogue. However, as Quaglio has observed, markers of vagueness are consistently higher in conversation as too much vague language leads to incomprehensibility on the part of the viewers. In the same way, the limited range of conversation topics and settings in televised dialogue is also found to affect the distribution patterns. He concludes that, on the whole, most of the linguistic features of naturally occurring conversation are shared by the sit-com corpus, thus making scripted speech a valuable substitute for spontaneous spoken data in foreign language classrooms (Quaglio 2009a: 149).

Within the scope of audiovisual translation, particularly film dubbing, researchers have investigated the formulaic nature of dubbed language. This has been described as a series of semantic and structural calques found to occur repeatedly across films, hence the term “translational routines” (Pavesi 2005: 48) and the sometimes derogatory term “dubbese”. There is a long list of routine translations, in both Italian and Spanish dubbed dialogue (see Pavesi (2005, 2008) on the English-Italian language pair and Romero Fresco (2006, 2009) on the English-Spanish pair), ranging from single items such as *absolutely*-‘assolutamente’, *sure*-‘sicuro’, to multiword units like *you know something?*-‘la sai una cosa?’, *forget about it*-‘scordatelo’, etc. Such expression patterns have been described as colloquial speech conveying the spontaneity of everyday language which, as a result of the translation process, tends to generate calques – e.g. vocatives and response forms in Italian dubbed language, as shown by Pavesi (2005, 2008, 2009); and discourse markers and intensifiers in the Spanish dubbed dialogue analysed by Romero Fresco (2009). According to the studies cited, therefore, it is the speech-like idiomatic quality of original scripted dialogue that most leaks into film translation, often making it sound unnatural and artificial. This perspective on translated film dialogue can be coupled with studies on the predictability of textual occurrences and frequencies as associated with particular scenes and scene types in a variety of original film dialogue. Specifically, Taylor (2004, 2006, 2008) describes the language of standardized exchanges such as service encounters, telephone conversations and other ritual moves involving greetings and leave-takings, and shows how a lot of the dialogue taking place on the screen is predictable. In the same way, in describing the features of scripted dialogue, Chaume (2001, 2004: 168) talks about "prefabricated orality", a concept that indicates the carefully planned and controlled nature of this kind of spoken discourse at each level of language, from prosodic to syntactic to lexico-semantic. The notion has been recently reprised by Baños-Piñero & Chaume (2009) in relation to audiovisual translation issues.

In conclusion, even though the concept of naturalness, as described in relation to spontaneous spoken data (cf. among others, Warren 2006), does not entirely pertain to onscreen scripted dialogue, the kinds of conversational resources that are found to accompany real-life interactions do, however, occur in filmic speech. These concern the interpersonal dimension of dialogue as well as the micro-level of discourse organisation and its cohesive potential – elements of hearer feedback including backchannels and non-minimal responses, speaker hesitations and repeats, interjections and discourse markers are all present in screen dialogue and their frequency and functions have been pointed out (see also Taylor 1999, 2008). Comprehensibility and reduction of processing effort on the part of the viewers as well as the overarching dramatic or humorous purposes of much film and TV dialogue can help explain the distributional differences observed when compared with unscripted speech. What is worth investigating further is the generalizability of these findings towards a reappraisal of film stylistics from a linguistic and phraseological perspective.[1] In order to do so, the process of enquiry is reversed. In fact, instead of taking available descriptions of the grammar and pragmatics of conversation to check one more time how they fit film dialogue and whether the same features are present and to what extent, the present investigation embarks on a frequency-driven analysis of filmic speech aimed at finding out the phraseological profile of the latter, i.e. at identifying frequent formulae to see how typical they are of the kind of orality represented by the sample under examination.

2. A phraseological approach: clusters

Predictions regarding systematic variations in language use in different communicative situations (what is commonly termed register variation) are supported by the phraseological approach. We are referring mainly to the theory of phraseology as developed by John Sinclair through the "idiom principle" (Sinclair 1991). That is to say, to the idea that the unit of meaning is no longer the single word. Rather, "words tend to go together and make meanings by their combinations" (Sinclair 2004a [1996]: 29). This tendency goes beyond simple collocations and idioms in the traditional sense, which is strictly associated with semantic opacity and non-compositionality, lexico-syntactic fixedness and institutionalisation. It is pervasive and concerns a wide range of multi-word combinations, as shown by corpus data. It is indeed corpus linguistic research with its systematic analysis of repeated usage in large sets of concordances that has brought to the fore the key role of phraseological expressions in language. As pointed out by Granger and Meunier in their (2008) introduction to an interdisciplinary volume on phraseology, before the advent of corpus linguistics, only the most fixed and opaque multi-word combinations offered themselves as an object of study. Now, however, a much wider variety of expressions displaying both lexico-syntactic variability and semantic compositionality are being investigated. Obviously, this has implications for both automated language processing and language pedagogy (see Sinclair (2004b) and the volume on phraseology in foreign language learning edited by Meunier and Granger in 2008). In fact, the new unit of meaning, the "lexical item", is the result of words repeatedly co-selecting each other in a corpus, with co-selection involving collocational and colligational patterns, semantic preference and semantic prosody (Sinclair 2004a [1998]). In Sinclair's latest contribution, this notion has been reconceptualised as "Meaning Shift Unit". An MSU is an extended unit of meaning corresponding to a shift in the ambient meaning (Sinclair in press). Although this is an abstract category, it is supported by frequency observations on recurrent chunks in a corpus variously called "clusters" (Scott 2004), "n-grams" (e.g. Fletcher 2007), "lexical bundles" (Biber et al. 1999), to name but a few of the terms used in the literature.

Clusters are defined as groups of words which are found repeatedly together in each other's company, in sequence.[2] They are "repeated strings which may have little or no psychological reality for speakers" (Scott and Tribble 2006: 19), but which "give insights into important aspects of the phraseology used by writers in specific contexts" (Scott and Tribble 2006: 132). Lexical bundles are frequent multi-word sequences which vary according to the register examined. Their structural and functional characteristics are found to be different in speech and writing, thus they provide a good indicator of register variation (Biber and Conrad 1999; Biber, Conrad & Cortes 2004). They are identified on the basis of their frequency in a "radically" corpus-driven fashion and therefore they are neither necessarily complete structural units nor perceptually salient (Biber 2009: 283). In sum, clusters or bundles are uninterrupted sequences of words found to occur repeatedly in a corpus. Also, in both accounts there is a focus on style and register variation: different styles and registers are characterised by different distributions of clusters, as these are indicative of "preferred ways of saying things" (cf. Altenberg 1998).[3] A frequency-driven approach like this goes well with a broad pragmatic concept of idiomaticity, which is typical of discourse-functional approaches (see Moon 1998; and more recently Corrigan et al. 2009). In the present paper, clusters as generated by a software for text analysis, namely *Wordsmith Tools version 4.0* (Scott 2004), offer the operational basis on which to ground an analysis of the phraseology of contemporary filmic speech. The corpus and the methodology used are illustrated in the next section.

3. Data and methodology

In order to investigate the formulaicity of contemporary filmic speech, two sample corpora have been used. One is the original component of the *Pavia Corpus of Film Dialogue*, a parallel corpus comprising both British and American films and their Italian dubbed versions. The other is a small addition compiled *ad hoc* by the author of this paper according to the same sampling criteria and conceived of as a further test case for the findings based on the former. The movies sampled in the PCFD at the time of writing cover the time span 1995–2004 and are all defined as "conversational" films, i.e. portraying interactions of different kinds, all characterised by face-to-face conversation in contemporary settings (on naturalistic drama see also the discussion in Richardson 2010). As a guarantee of their status and diffusion within the cultures that produced and consumed them, they were successful both with the critics and the general public (Freddi & Pavesi 2009). The sampling cuts across rigid film genre distinctions, ranging from the social cinematography of Ken Loach and Mike Leigh to romantic comedies such as *Notting Hill* and *Bend it Like Beckham*. There are no costume films nor musicals, but action movies such as *Ocean's Eleven* and thrillers like *One Hour Photo* are included. The second dataset consists of four extra movies chosen according to the same sampling criteria as above: two American (*Lost in Translation* and *Michael Clayton*) and two British (*Snatch* and *Love Actually*). The time span covered is roughly the same except for the slight update represented by *Michael Clayton*, and the genres match those in the PCFD.

Both *Lost in Translation* and *Love Actually* are romantic comedies, like, for example, *Notting Hill* and *Sliding Doors* are in the PCFD; *Snatch* is a heist film, very much like its PCFD American counterpart represented by the *Ocean's* saga; while *Michael Clayton* is a lawyer movie like *Erin Brockovich* in the other corpus. At the time of writing, the PCFD runs about 23 hours of film in each language (there are 12 films altogether) totalling almost 118,000 words of original dialogue and slightly fewer words in the translations-adaptations (almost 112,000). The second dataset consists of only 4 films, the total size of the corpus being roughly 60,000 words between the two languages, thus adding almost another 8 hours of film for each version.[4]

For the purposes of the present study, the translational component of each corpus is used only as background data, as issues of film translation go beyond our scope. It should also be noted that the text in the corpus is the orthographic transcription of the actual lines uttered by the characters on screen. In other words, the version included in the corpus differs greatly from the available scripts in terms of features of spokenness often added by actors when they are shooting, such as hesitations, repeats, false starts and interruptions, overlaps, etc. (see on this Taylor 2004: 79–80). Since the speech-like and colloquial nature of conversation in scripted dialogue is of interest to the present investigation, only a careful transcription of the final product could be used as the object of study. We do not know as yet which features of conversation will be captured by the clusters list, but, as they affect the frequency counts, they might be part of some repeated sequences relevant to the phraseological profile of the corpus, so a different picture would emerge if they were left out.

The two corpora are processed using the concordance software for text analysis *Wordsmith Tools v. 4.0*, which is also used to create clusters lists. The initial analysis is therefore frequency-driven in that it is the clusters rather than previous accounts of the grammar of conversation that offer an initial description of the sample of filmic speech in question (differently from the literature expounded in Section 1). The same procedure is adopted for both the PCFD and the additional dataset to see if results corroborate each other. The formulae thus identified are subsequently compared with available reference corpora of spoken British and American English (the spoken components of the British National Corpus and the Corpus of Contemporary American English) in order to explore differences and similarities in the distributions. Finally, the functions of the various formulae are identified and discussed also in relation to their being present in the large general corpora of spoken British and American English. It is claimed that a data-driven approach points to functional differences and register-specific usage.

4. Film phrases: distributions and functions

4.1. *4-word clusters in film dialogue*

Figure 1 offers a graphic representation of the top 25 frequencies in the PCFD. This allows the visualization of the flattening out of the frequency distribution with the curve becoming asymptotic. For this reason, a cut-off point is established on the eleventh cluster in the list.

In Table 1, a list is given of the top 11 clusters with raw frequency, normalized frequency per hundred thousand words, and the number of films (texts) in which they are found to occur. The last information, which is displayed in the right-most column, gives a clue as to how clusters are distributed across the films in the corpus – sometimes called "dispersion" (Scott and Tribble 2006: 29) – and should be kept in mind for the comparative phase of the discussion. At this stage, it should be noticed that the 11 most common clusters in the PCFD contain the pronouns *you* or *I*, or both, which are the most frequent words in the corpus. This comes as no surprise given the dialogic nature of the text-type under consideration and is in line with Quaglio's findings on two of the most frequent lexical bundles in the sit-com *Friends*, namely *I can't believe you* and *Thank you so much*, both containing *I* (elided in thanks) and *you.* According to Quaglio, this is

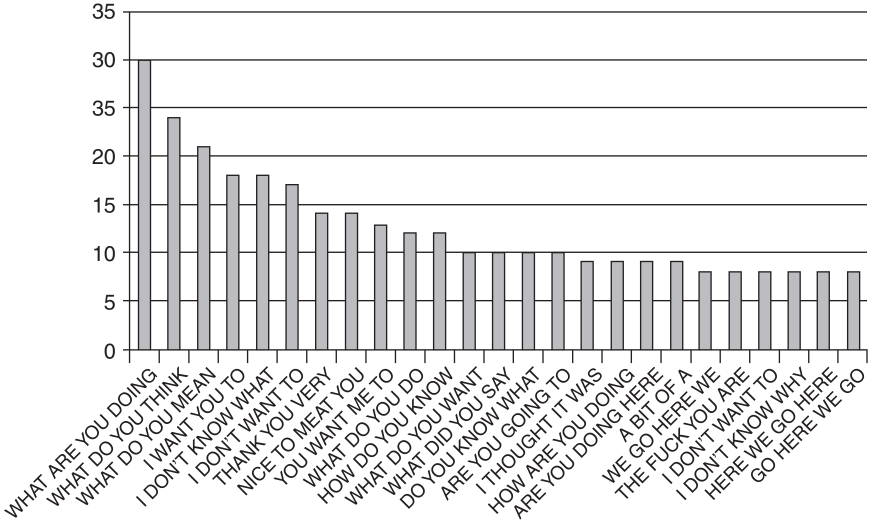

Figure 1. *Frequency of 4-word clusters in the PCFD*

Table 1. *Top 11 clusters in the PCFD*

N	*Cluster*	*Freq.*	*PHT*	*Texts*
1	WHAT ARE YOU DOING	30	25	11
2	WHAT DO YOU THINK	24	20	9
3	WHAT DO YOU MEAN	21	18	11
4	I WANT YOU TO	18	15	9
5	I DON'T KNOW WHAT	18	15	6
6	I DON'T WANT TO	17	14	7
7	THANK YOU VERY MUCH	14	12	8
8	NICE TO MEET YOU	14	12	6
9	YOU WANT ME TO	13	11	8
10	WHAT DO YOU DO	12	10	6
11	HOW DO YOU KNOW	12	10	6

evidence for the involved nature of the exchange and is even more so when compared to other kinds of conversational data (Quaglio 2009a: 100).

Why 4-word clusters? There is indeed a problem with the horizon of the multi-word sequences which are searched for in a corpus. Obviously, the longer the span we set, the fewer the clusters we obtain. In previous research, 3-word clusters have been looked at (Freddi in press), which contained fragments of *what*-questions (see the top 3 and no. 10 in Table 1). They displayed the interrogative pronoun followed by subject-verb inversion or a potentially intervening swearword (e.g. WHAT ARE YOU, WHAT DO YOU, WHAT THE HELL, etc.). However, 3-word clusters come up with a lot of overlap, which is hard to filter out (the list contained sequences such as WHAT ARE YOU and ARE YOU DOING, or WHAT DO YOU and DO YOU MEAN, DO YOU THINK, DO YOU WANT, etc.). Therefore, 4-word clusters have been chosen here in order to avoid the overlap of the 3-word clusters and to obtain a higher number of independent chunks. Also, a string of 4 words in a cluster makes it comparable to previous research on clusters and register variation done by Biber et al. (1999) and Biber and Conrad (1999). See the discussion in Section 4.2.

The same search was done on the smaller dataset; the same clusters appear on top of the list although there is a skew effect observable in the frequencies due to the presence of one film in particular, *Love Actually*, which is longer than the others and has songs containing refrains. This brings us back to data sampling and transcription practices. On the one hand, an excessively small sample is more sensitive to the presence of a single film. On the other, songs could be left out in principle, but it was decided not to do so when they constituted an inherent part of the plot and were linked to the characters' lines, as in this case. The data has been ordered according to dispersion (i.e. the number of films in which the clusters were found) to avoid this problem. Notice the steep decrease: only the top 3 clusters are found in at least 75% of the corpus (Table 2). However, the interpersonal routine *Nice to meet you* (rank no. 4), already found as prominent in the PCFD, and other *what*-questions (nos. 5, 7) or chunks thereof (no. 11) are also in the list. With rank no. 12, clusters come from just the one film mentioned above and are no longer worth considering.

Table 2. *Top clusters in test dataset sorted according to dispersion*

N	*Cluster*	*Freq.*	*%*	*Texts*	*%*
1	WHAT ARE YOU DOING	8	0,02	4	100,00
2	WHAT DO YOU THINK	7	0,02	3	75,00
3	YOU WANT ME TO	5	0,02	3	75,00
4	NICE TO MEET YOU	10	0,03	2	50,00
5	WHAT DO YOU WANT	8	0,02	2	50,00
6	I DON'T KNOW WHAT	7	0,02	2	50,00
7	WHAT DO YOU MEAN	7	0,02	2	50,00
8	ARE YOU DOING HERE	6	0,02	2	50,00
9	I CAN TELL YOU	5	0,02	2	50,00
10	I DON'T KNOW I	5	0,02	2	50,00
11	WHAT THE HELL IS	5	0,02	2	50,00
12	FEEL IT IN MY	11	0,03	1	25,00
13	I FEEL IT IN	8	0,02	1	25,00
14	ALL I WANT FOR	7	0,02	1	25,00
15	I WANT FOR CHRISTMAS	7	0,02	1	25,00
16	IT IN MY TOES	6	0,02	1	25,00
17	LOVE LOVE LOVE LOVE	6	0,02	1	25,00

4.2. *Film clusters and the BNC*

In this section, the results of the comparisons drawn between the film corpus and the BNC are presented and discussed. Firstly, the clusters from the PCFD are compared to the BNC spoken, particularly face-to-face spontaneous conversations. Secondly, the BNC spoken without any genre restriction is used. The results are displayed in the following histograms in order to facilitate the visualisation of the distributions. However, to avoid misinterpreting the results, a word of caution is in order concerning the text-types sampled in the spoken BNC.

The corpus contains approximately 100 million words of text, 90% written and 10% spoken, divided into various sub-genres. The spoken component includes 10 million words from 1970 to 1994 between demographically sampled and context-governed data (respectively 4 and 6 milion words). The corpus is a closed sample, i.e. not updated with fresh material. The demographic sampling reflects the sociolinguistic variation of the speakers according to age, sex, occupation and geographical provenance within the UK, while context-governed sampling identifies the different kinds of interactions taking place, including business meetings, trade unions meetings, academic lectures, TV or radio news, political speeches, parliamentary proceedings, radio phone-ins, etc. A comparative view of 4-word clusters in the three corpora is given in Figure 2 below.

The frequency of the top 4 clusters in the PCFD (the black bar) is significantly different from the BNC distribution, both for the face-to-face conversations (the white bar) and for the overall spoken corpus (the grey bar). The difference is such that even without calculating a test for statistical significance, we can hypothesise that register-specificity and functional specialisation explain the high frequency of the top 4 clusters (namely, the *what*-questions and the directive *I want you to*). It should also be remembered that they have been found to occur in almost all films throughout the PCFD as well as in the additional dataset. Particularly, the very high frequency of the top cluster *what are you doing* can be explained by recalling the principle of conflict as constitutive of screenwriting (cf. among others Aimeri 2007; Seger 2009). Characters take action on screen and their actions are challenged by other characters who will therefore address them

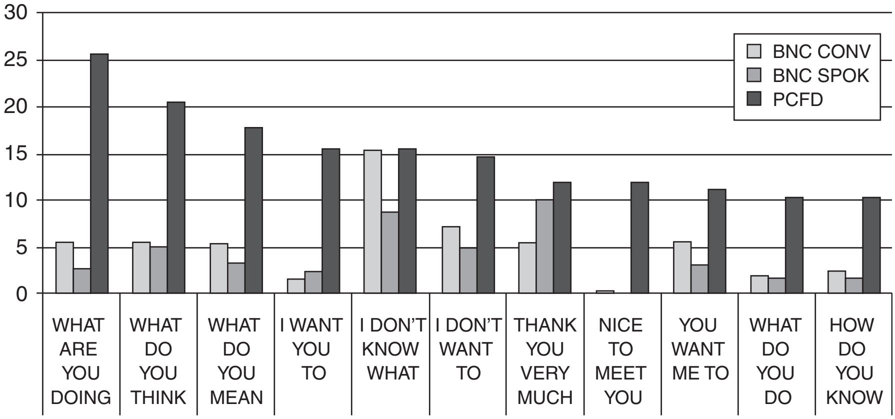

Figure 2. *4-word clusters: comparative view 1*

directly and ask *what are you doing?* or *what are you doing here?* These are very common in the film corpus, as is visible from the concordance lines (cf. Section 4.4). The question becomes an almost fixed formula whose purpose is not only the exchange of information, as in real-life situations, but also what can be termed, following Kozloff (2000: 33), the "anchorage of the diegesis", i.e. it is part of the main narrative. In contrast, in the kinds of conversations sampled in the BNC, the people are less likely to be doing something, possibly because of the way the data are recorded. In fact, the speakers are carrying around a tape-recorder, without, actually, doing much (e.g. in informal conversations, chatting and after-dinner talking). This is also because of the types of interactions, especially in the context-governed part of the corpus, as the smaller proportion of the grey bar shows.

The clusters *I don't know what* and *I don't want to*, occupying ranks 5 and 6 respectively in the PCFD, are much closer to the BNC frequencies, especially when compared to conversation. Interestingly, 5 and 6 not only fit the general pattern of a large number of bundles in conversation, which, as stated by Biber et al. (1999: 991), "are constructed from a pronominal subject followed by a verb phrase plus the start of a complement" (cf. the clusters *I don't know what*, *I don't know how*, *I don't want to*, *I don't think so*, etc. from the *Longman Corpus of Conversation*), but are the two most common 4-word clusters found by Biber and Conrad (1999: 183) and Biber et al. (1999: 994) to be typical of conversation when compared to written academic prose. In Biber *et al.*'s corpus they both occur over 100 times per million words (i.e. over 10 times PHT), which again is comparable to the BNC conversations frequencies in Figure 2. The function of both clusters is to introduce the speaker's stance relative to the information that follows, which is perhaps why their use is as widespread across all kinds of conversational data considered, whether scripted or unscripted.

Very little has been observed to be typical of filmic speech in the remaining clusters. An exception is presented by no. 8 *nice to meet you*, which was found to occur only 5 times in the overall spoken BNC, 4 of which occurred in the face-to-face conversations, and also possibly no. 9, where the directive *you want me to* is part of a larger interrogative sequence and is symmetrical to the other directive *I want you to* in rank no. 4. The high incidence of the greeting formula *nice to meet you* is probably due to the different kinds of situations portrayed in films. People are introduced to each other for the first time much more often in films than in the kind of real-life conversations sampled in the reference corpora. Quaglio (2009a: 35) reports the same, noticing how greeting exchanges are extremely frequent in the *Friends* corpus because of the structure of the scenes, which often commence with characters arriving in places and meeting one another (although not for the first time). It should also be noted that clusters in ranks 8, 10 and 11 were only found in 50% of the movies in the corpus. Therefore, whether they are typical or not can hardly be assessed in relation to a reference corpus.

4.3. *Film clusters and the COCA*

The data can also be compared to the COCA, given that half the movies in the film corpus are American productions. Again, knowledge of the sampling criteria followed by the COCA compiler when choosing the texts to include in the corpus is necessary to better interpret the results of the comparisons.

The corpus contains more than 410 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts (five sections). It includes 20 million words each year from 1990 to 2010 and it is updated once or twice a year (the most recent texts are from Summer 2010). However, the sampling of the spoken component differs from the BNC. The spoken COCA comprises approximately 87 million words, it contains only transcripts of conversations from TV and radio programmes, such as, *All Things Considered*, *Newshour*, *Good Morning America* (broadcast by ABC), *The Today Show* (NBC), *60 Minutes* (CBS), *The Larry King Show* (CNN), etc. Leaving aside a discussion of whether this kind of data is in fact unscripted and representative of non-media varieties of spoken American English, its nature should be kept in mind when drawing the comparisons. Conversely and unlike the BNC, the fiction section in the COCA contains a relevant proportion of movie scripts (almost 12 million words). According to the compiler of the COCA, since movies are written by a screenwriter, they do not represent actual, unscripted speech and therefore are considered written, not spoken data (cf. the COCA site). In addition, and related to this, characters' names are not separated out from the lines, so part of the 12 million words of the movies subsection includes that information as well. This is different from how we calculated the total number of running words in the PCFD and in the test dataset for that matter, where we only took the lines leaving out the characters and setting information. This rules out a comparison between the PCFD on the one hand and the COCA movie sub-section on the other. The difference in size between the two samples is indeed too big – 117,000 tokens vs. 11,700,000 words – , that is to say the COCA movie sub-section is a hundred times bigger than the PCFD and even with the small addition of the 4 films from the test dataset, things would not change significantly.

So, the comparison that is worth drawing seems to be the one between the overall spoken section of the COCA (ca. 87 million words) and the movie sub-section in the COCA (ca. 11,700,000 words) by checking the frequency distributions of the clusters extracted from the PCFD. The results are shown in the next figure (see Figure 3).

The comparison shows that clusters nos 1 (*what are you doing*), 4, 9 (the directive chunks *I want you to* and *you want me to*) and possibly no. 3 (*what do you mean*) can be considered typical of movies (represented by the black bar), while the same cannot be said of no. 2 *what do you think* for which the frequencies are higher in the spoken section (the white bar) than in the movies. This might be due to the kinds of texts that are sampled in the COCA spoken section, mainly TV and radio interviews in which interviewers ask interviewees about their opinions on the various issues debated (*What do you think about/of…?*), and seek further explanation (*what do you mean…?*). The clarifying function is more distinctive of the types of interactions sampled in the spoken COCA and less typical of movies.

Furthermore, clusters nos 5 and 6 are as frequent in the overall spoken section as in the movies, replicating the same situation as with the BNC-PCFD comparison (see Figure 2). A slightly different trend from the BNC-PCFD is observable in cluster no. 7 *thank you very much*. This is practically absent from the COCA movie sub-section and significantly higher than the other clusters in natural spoken data. In contrast, when compared to the overall spoken BNC, *thank you very much* was higher in the PCFD, although not

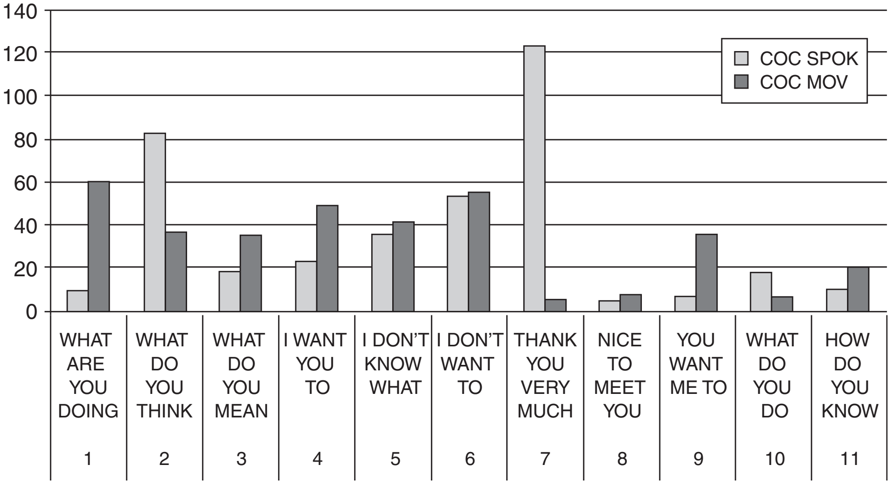

Figure 3. *4-word clusters: comparative view 2*

significantly so. Again, the very high frequency of this routine in the COCA spoken section might be linked to the recurrence of certain set situations and communicative moves in the data sampled in the corpus, like when the host of a TV or radio programme thanks his/her guest interviewee. Additionally, no. 8, the greeting formula, has very few occurrences in both the spoken COCA and in the movies (unlike the PCFD). No. 9, the directive *you want me to*, is higher in the movies than in the spoken data, very much like the previous comparison between the PCFD and the BNC and thus adding evidence to its being characteristic of film dialogue. Nos. 10 and 11 are not considered relevant because of their sparsity in the original data, namely in the PCFD, where they were only associated with half or fewer than half of the films in the corpus.

4.4. *The functions of film phrases: film stylistics revisited*

If we check the concordances for each cluster in order to observe them in context, we see that there are sequences of two kinds. On the one hand, *what*-questions like *What are you doing?* and *What are you doing here?* are found throughout the film corpus. On the other hand, fixed expressions with a specific conversational-interpersonal function also seem to be quantitatively prominent and diffuse, see, for example, *Thank you very much*, *Nice to meet you*, *How are you doing?*

The first group is of particular interest as it comprises often confrontational questions, signalling that some kind of conflict is going on between the two characters involved in the exchange. Sometimes they are "enriched" by expletives, a mark of emotionally-loaded language, as in *What the hell/the fuck are you doing here?*. Notably, a problem with extracting clusters in the way we have done here, that is on the basis of the frequency of contiguous word-forms in a given span, is that they do not show the amount of lexical variation within the unchanging structural combination (see also Biber 2009). This is why the concordance helps to see sequences interrupted by a swearword and variations on the interrogative patterns *What do you think?* and *What do you mean?* like, for example, *What do you want?* (also frequent in the film corpus, but less so, see Figure 1). The same can be said of *What are you talking about?*, which indeed counts as a 5-word cluster, and other questions like *What's going on?*, *What is it with you?*, *What's wrong with you?*, etc., which are not frequent enough to count as cluster if considered individually, but taken together express an attitude of conflict.[5] This is in line with Van Lancker-Sidtis and Rallon's observation that conventional or formulaic expressions "are generally laced with attitudinal or emotional innuendos" (2004: 208). Besides, they maintain their prototypical function and explicitating purpose of prompting the missing bit of information, which is functional to the advancement of the plot. In the first example, Jamie (played by British actor Colin Firth) gets back home where he left his malingering wife, finds his brother unexpectedly there and utters the line *What the hell are you doing here?*, expressing surprise if not conflict. The question sets off a scene which ends in Jamie's finding out that his wife is cheating on him:

(1) from *Love Actually*
Jamie is back home
JAMIE Hello! What the hell are you doing here?
JAMIE'S BROTHER Oh, I just popped over to borrow some old CDs.
JAMIE The lady of the house let you in, did she?
JAMIE'S BROTHER Yeah.
JAMIE Lovely, obliging girl.
JAMIE'S BROTHER Yeah.
JAMIE I-I just thought I'd pop back from the reception to see if she's better.
JAMIE'S BROTHER Oh.
JAMIE Listen, erm, I've been thinking. I think perhaps we ought to take mum out for her birthday on Friday. What do you think? I just feel we've been bad sons this year.
JAMIE'S BROTHER Okay, sounds fine. A bit boring but fine.
JAMIE'S WIFE (VOICE) Hurry up big boy! I'm naked and I want you at least twice before Jamie gets home!

More examples of the sequence *What are you doing?* from the PCFD are given in Table 3. It is interesting to note the use of the vocative *man* and the tag *eh* both reinforcing the interpersonal dimension of the interrogative (lines 33 and 136, both from *Billy Elliot*) and the insult accompanying the question in *Ae Fond Kiss* (line 233). The translations have been left in, as they explicitate the fixed and argumentlike quality of the situation through the marked verb 'combinare' instead of the unmarked 'fare'-*do*, and the addition of the weak connective 'ma' (Eng. *but*) in turn-initial position. These questions very often do not even expect an answer. As they signal conflict or at least tension between two characters in the flow of conversation, they are emptied of their actual propositional content, i.e. asking for information, and become a sort of speech act of criticising or expressing

Table 3. *Concordances of What are you doing? from the PCFD*

Film	*Line No.*	*Language*	*Character*	*Text*	*Translator*
Ae Fond Kiss	105	English	TAHARA	What are you doing?	
Ae Fond Kiss	105	Italian	TAHARA	Cosa stai combinando?	Depaolis Federica
Ae Fond Kiss	233	English	DANNY	What are you doing? You stupid bastard.	
Ae Fond Kiss	233	Italian	DANNY	Ma guarda che hai combinato!	Depaolis Federica
Ae Fond Kiss	278	English	HAMMID	What are you doing later on? No, let's go!	
Ae Fond Kiss	278	Italian	HAMMID	Che cosa fate dopo?	Depaolis Federica
Ae Fond Kiss	287	English	CASIM	What are you doing here?	
Ae Fond Kiss	287	Italian	CASIM	Che ci fai tu qui?	Depaolis Federica
Ae Fond Kiss	300	English	CASIM	What are you doing?	
Ae Fond Kiss	300	Italian	CASIM	Che cosa stai facendo?	Depaolis Federica
Bend it like Beckham	29	English	FRIEND 1	Yeah. What are you doing here man? You haven't left everything to the last minute man, have you?	
Bend it like Beckham	29	Italian	FRIEND 1	Ciao, sposina! Che ci fai qui, non avrai rimandato tutto all'ultimo momento, vero?	Caporello Elettra
Bend it like Beckham	846	English	JULES	What are you doing?	
Bend it like Beckham	846	Italian	JULES	Ma che state facendo?	Caporello Elettra

Bend it like Beckham	900	English	JESS	Tony, what are you doing?!	
Bend it like Beckham	900	Italian	JESS	Tony! Tony, ma che vuoi fare?	Caporello Elettra
Bend it like Beckham	1004	English	JESS	What are you doing here?	
Bend it like Beckham	1004	Italian	JESS	Come mai sei qui?	Caporello Elettra
Billy Elliot	33	English	GEORGE WATSON	(…) This is man-to-man combat, not a bloody tea dance. What are you doing, man? Hit him! He’s just pi	
Billy Elliot	33	Italian	GEORGE WATSON	(…) È un combattimento corpo a corpo, non un tè danzante! Dove accidenti vai? Colpiscilo! Chris ti prende	Cosolo Carlo
Billy Elliot	87	English	DAD	What are you doing going around like creeping Jesus?	
Billy Elliot	87	Italian	DAD	Che stai facendo che ti aggiri come un ladro?	Cosolo Carlo
Billy Elliot	91	English	DAD	What are you doing?	
Billy Elliot	91	Italian	DAD	Che stai facendo?	Cosolo Carlo
Billy Elliot	136	English	TONY	Got enough food there, scab? What are you doing, eh?	
Billy Elliot	136	Italian	TONY	Hai preso abbastanza da mangiare, crumiro? Che stai facendo? eh?	Cosolo Carlo
Billy Elliot	251	English	BILLY	What are you doing?	
Billy Elliot	251	Italian	BILLY	Che stai facendo?	Cosolo Carlo
Billy Elliot	445	English	BILLY	What are you doing?	

(*Continued*)

Table 3. *Continued*

Film	*Line No.*	*Language*	*Character*	*Text*	*Translator*
Billy Elliot	445	Italian	BILLY	Cosa stai facendo?	Cosolo Carlo
Crash	26	English	KIM LEE	(…) You fuck you too! Fuck you too! So, what are you doing, are you gonna arrest me?	
Crash	26	Italian	KIM LEE	(…) Lei fatto incidente e lei ola paga pe' lipalale mia macchina! Lei no amelicana! Lei mexcana! Ffanculo Mexico! Ffanculo somblelo!	Ottoni Filippo
Dead Man Walking	344	English	SISTER HELEN	(…) So he's on his deathbed and a friend comes to visit and he sees him reading the Bible. The friend says "W.C., you don't believe in God, what are you doing reading the Bible?" and Fields says	
Dead Man Walking	344	Italian	SISTER HELEN	(…) Insomma sta sul letto di morte, un amico va a trovarlo e vede che legge la Bibbia, allora gli chiede: "W.C. tu non credi in Dio, ma che fai leggi la Bibbia?" E Fields dice (…)	Bertini Lorena
Dead Man Walking	834	English	MATTHEW PONCELET	Hey, you know what I'm doing. What are you doing?	
Dead Man Walking	834	Italian	MATTHEW PONCELET	Lo sai. Tu che stai facendo?	Bertini Lorena
Erin Brockovich	168	English	ERIN	What are you doing making all that goddamn noise?	
Erin Brockovich	168	Italian	ERIN	Come vi salta in mente di fare tutto questo casino?	Mete Marco

Erin Brockovich	416	English	ERIN	Ooh… It's some slim pickings here, baby. Maybe that's Ed McMahon. Let's go see. Wow. Oh… wrong Ed. What are you doing here?	
Erin Brockovich	416	Italian	ERIN	E con che le pago? Oh… Qui la vedo magra, piccina mia… Ah… Potrebbe essere Eddie Murphy, andiamo a vedere… Ah, l'Eddie sbagliato. Che ci fai qui?	Mete Marco
Finding Forrester	404	English	JAMAL	What are you doing?	
Finding Forrester	404	Italian	JAMAL	Ma che sta facendo?	Caporello Elettra
Notting Hill	364	English	BELLA	Quickly, quickly, quickly. Talk very, very quickly. What what are you doing here with Anna Scott?	
Notting Hill	364	Italian	BELLA	Presto, presto, presto. Parla più in fretta che puoi. Che cosa ci fai qui con Anna Scott?	Vairano Francesco
Ocean's Eleven	95	English	BARRY	I'm out what are you doing?	
Ocean's Eleven	95	Italian	BARRY	Ma che fai?	Mete Marco
Ocean's Eleven	593	English	TESS	What are you doing here?	
Ocean's Eleven	593	Italian	TESS	Che ci fai qui?	Mete Marco
Ocean's Eleven	644	English	MR BENEDICT	What are you doing?	
Ocean's Eleven	644	Italian	MR BENEDICT	E che cosa fa?	Mete Marco

disagreement. This function is also visible in *What are you doing making all that goddamn noise?* (line 168 from *Erin Brockovich*, but also 87 from *Billy Elliot*) where the construct with the gerund adds to the function of the question. Clearly, the meaning of the question is not always confrontational, as, for example, in line 278 from *Ae Fond Kiss*, where it is a real question expecting an answer (cf. *What are you doing later on?*).

The same *what*-questions can be searched for in the spoken components of the reference corpora used in the previous Sections (cf. 4.2 and 4.3) to check their distribution and functions therein. To illustrate this, let us take the string WHAT DO YOU MEAN in the spoken BNC. The search returned 336 hits in 165 different texts (with a frequency of 32.28 instances per million words, cf. also Figure 2). The following examples highlight its use in dialogic contexts to frame the interlocutor's own words on which some clarification is asked for. See example (2) from a radio broadcast phone-in:

(2) HUV 72
What do you mean you tried for two days?

In the extracts in (3), taken from demographically sampled conversations, the framing function is made explicit by the transcribers' choice to separate the reporting from the reported part with a comma:

(3) KB7 4064
What do you mean, left the table?
KB7 4451
What do you mean, been conning you?
KB7 6063
What do you mean, nothing?

Interestingly, the last concordance line in example (3) was also found in the film corpus, where the notation used to indicate a piece of reported speech is the double quotes. The meaning of the question could be glossed as "taking up the interlocutor's words to either rebut or challenge them", cf. (4) from the film *Finding Forrester*:

(4) TERREL What do you mean "nothing"? This getting in the way of your plans or something?

If we look at the list of collocates of WHAT DO YOU MEAN in the spoken component of the BNC, the most statistically significant collocate is the question mark, as in *What do you mean?* making up a whole turn, which is also frequent in the two film corpora. On the other hand, the second most frequent collocate to the right in BNC spoken is the preposition *by* (as in *what do you mean by…*) whose use, however, seems to be restricted to institutionalised discourse types such as lectures, business meetings and the like. There is no such usage in the film corpus, showing that it belongs to more formal contexts.

The second group of routine expressions represents the verbal accompaniment of highly ritualised acts of daily experience. Their relevance to film dialogue vs. natural conversation has already been discussed in relation to what Quaglio (2009a: 135) has termed "movement", that is the characters walking into a place and exchanging greetings with other characters like in ordinary encounters. This, in particular, explains the high incidence of *Nice to meet you* and *How are you doing?* in the films (cf. Figure 1). Following Van Lancker-Sidtis and Rallon's categorisation, both groups would be classified as speech formulae, that is "the expressions used in conversational interaction" (2004: 209). Among the examples the authors give are *See you later!*, *Let's call it a day* and *You don't say!* and in their analysis of conversational dialogue in the screenplay *Some Like it Hot* speech formulae constitute by far the highest number of formulaic expression types (and highest number of repetitions per type). Although their analysis is not frequency-driven, many of the speech phrases they find in the screenplay are similar to our findings: thanks and apologies are frequent, and a variety of *what*-questions occur repeatedly (including *What's the matter?*, *What's going on here?*, *What do you think you are doing?*, *What are you talking about?*, etc.). Also, the authors draw attention to the fact that formulaic expressions often occur in a context of confrontation and disagreement and concede that disagreement is part of the texture of the screenplay they chose to analyse (2004: 220).

In conclusion, frequent phrases work on a double level:

i) *diegetic*, i.e. internal to the fictional world of narrated situations and events. They have a plot-advancing function and contribute to the representation of conflict,

and

ii) *conversational-interpersonal*, i.e. mimicking the spontaneity of natural conversation as linked to the interpersonal sphere of dialogic interactions.

This double function can be linked to two fundamental purposes of film dialogue as formulated by Kozloff in her (2000) study of dialogue in film.

These are the anchorage of the diegesis and characters and the "verbal wallpaper" of ordinary conversational activities (Kozloff 2000: 47). The frequent *what*-questions identified under i) are verbal acts whose purpose is the anchoring of the narrative; they help to carry on the action while signalling some sort of challenge between two characters in the flow of conversation. The conversational routines in ii) are part of the mannerism of film as a realistic construct.

5. Conclusions

The description proposed in the present study on the phraseology of contemporary filmic speech is first and foremost frequency-based in that recurrent sequences of words in a film corpus (PCFD) have served as the starting point for the research. Furthermore, in order to make a comparison between spontaneous conversation and film dialogue and look into register-specific usage, the clusters thus identified have been compared to general spoken corpora, namely the spoken components of the BNC and the COCA. The comparisons have identified very few clusters which are typical of the register under examination, i.e. scripted film dialogue, most of them being common to natural spoken data. As for the methodological issues involved in using clusters as a tool for identifying phrases of film dialogue, one should note that clusters do not allow for variations on a pattern. In order to overcome this problem, the distributional patterns have been further checked against the concordances from the PCFD. The analysis has foregrounded chunks of speech functioning on two levels, namely diegetic and conversational. The former has been explained as triggering off the advancement of the situations and events internal to the fictional world narrated by film (the interrogative pattern); in this respect it reflects the register-specificity of formulae in film dialogue. The latter has been understood as functioning in real life conversation mainly interpersonally, therefore, it is said of formulae which are mimetic of natural conversation in English. This takes us back to film theory and the development of both the diegetic and mimetic dimensions of film, which works simultaneously as a telling, a narration or verbal activity, and a showing, a mimesis. Dialogue, as one of the codes in the complex semiotic environment of film, has been shown to function both narrationally and mimetically, also in virtue of its formulaicity. In this regard, the study corroborates previous observations on the predictability of filmic speech as expressed by Taylor (2004, 2008) and resonates with Eikhenbaum's concept of "divination" on the part of the film viewer:

> Cinema demands of the viewer a certain special technique for divination, and this technique will of course become more complex as the art of film-making develops. Directors already make frequent use of symbols and metaphors, the meaning of which depends directly on current verbal metaphors. Film viewing is accompanied by a continual process of internal speech. We have already grown accustomed to a whole series of typical patterns of film-language; the smallest innovation in this sphere strikes us no less forcibly than the appearance of a new word in language. To treat film as an absolutely non-verbal art is impossible.
>
> (Eikhenbaum 1974: 14)

In conclusion, the phraseological approach used here allows a reconsideration of film stylistics, a reconsideration which focuses on the language used in films and highlights its role as integral to the other semiotic codes (esp. images and montage).

Notes

Correspondence address: maria.freddi@unipv.it

1. I use "film stylistics" as a homage to Eikhenbaum's (1974) seminal paper on film as a form of art.
2. In the *Wordsmith Tools* handbook we read "each n-word cluster will be stored, if it reaches n words in length, up to a punctuation boundary, marked by semicolon, comma, full-stop, exclamation and question marks".
3. For a discussion of recurrent word-combinations as "preferred ways of saying things" as in Altenberg (1998), see Granger and Paquot (2008); Corrigan et al. (2009: xiv) talk about restricted distributions and formulae as "true hallmarks of style". On phraseology and style, see also some of the contributions in the two volumes by Burger et al. (2007).
4. For details concerning each film, runtime and number of running words, see Freddi and Pavesi (2009: 99) and Freddi (in press).
5. Contracted forms like *what's* as opposed to *what is* do of course represent a problem for the automatic computation of clusters but because contractions are frequent in conversation, they should be taken into account when analyzing clusters.

References

Aimeri, Luca. 2007. *Manuale di sceneggiatura cinematografica: Teoria e pratica*. Torino: Utet.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In Anthony P. Cowie (ed.), *Phraseology*, 101–122. Oxford: Oxford University Press.

Baños-Piñero, Rocio & Frederic Chaume. 2009. *Prefabricated orality: A challenge in audiovisual translation. inTRAlinea Special Issue on The Translation of Dialects in Multimedia*. Online at: http://www.intralinea.it/specials/dialectrans/eng_more.php?id=761_0_49_0_M (accessed 30 April 2011).

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multiword patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3). 275–311.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Biber, Douglas & Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In Hilde Hasselgard & Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 181–190. Amsterdam/Atlanta: Rodopi.

Biber, Douglas, Susan Conrad & Vivian Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics.* 25(3). 371–405.

Burger, Harald, Dobrovol'skij Dmitrij, Peter Kühn & Neal R. Norrick (eds.). 2007a. *Phraseology: An international handbook of contemporary research. Vol. 1.* Berlin: de Gruyter.

Burger, Harald, Dobrovol'skij Dmitrij, Peter Kühn & Neal R. Norrick (eds.). 2007b. *Phraseology: An international handbook of contemporary research. Vol. 2*. Berlin: de Gruyter.

Chaume, Frederic. 2001. La pretendida oralidad de los textos audiovisuales y sus implicaciones en traduccion. In Rosa Agost & Frederic Chaume (eds.), *La Traducción en los Medios Audiovisuales*, 77–87. Castelló de la Plana: Publicacions de la Universitat Jaume I.

Chaume, Frederic. 2004. *Cine y Traduccion*. Madrid: Catedra.

Corrigan, Roberta, Edith A. Moravcsik, Hamid Ouali & Kathleen M. Wheatly. 2009. Introduction: Approaches to the study of formulae. In *Formulaic language. Vol. 1 Distribution and historical change,* xi–xxiv. Amsterdam/Philadelphia: Benjamins.

Eikhenbaum, Boris. 1974. Problems of film stylistics. *Screen* 15(3). 7–34.

Fletcher, William. 2007. *kfNgram*. Online at: http://kwicfinder.com/kfNgram/kfNgramHelp.html (accessed 30 April 2011).

Freddi, Maria. in press. *What AVT can make of corpora: Some findings from the Pavia Corpus of Film Dialogue*. In Mary Carroll, Pilar Orero & Aline Remael (eds.), *Media for all: Quality made to measure, (Approaches to Translation Studies.)* Amsterdam: Rodopi.

Freddi, Maria & Maria Pavesi (eds.). 2009. *Analysing audiovisual dialogue: Linguistic and translational insights.* Bologna: Clueb.

Granger, Sylviane & Fanny Meunieur. 2008. Introduction: The many faces of phraseology. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective,* xix–xxviii. Amsterdam/Philadelphia: Benjamins.

Granger, Sylviane & Magali Paquot. 2008. Disentangling the phraseological web. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 27–49. Amsterdam/Philadelphia: Benjamins.

Kozloff, Sarah. 2000. *Overhearing film dialogue*. Berkeley/Los Angeles: University of California Press.

Meunier, Fanny & Sylviane Granger (eds.). 2008. *Phraseology in foreign language learning and teaching*. Amsterdam/Philadelphia: Benjamins.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon.

Pavesi, Maria. 2005. *La traduzione filmica: Aspetti del parlato doppiato dall'inglese all'italiano.* Roma: Carocci.

Pavesi, Maria. 2008. Spoken language in film dubbing: Target language norms, interference and translational routines. In Delia Chiaro, Christine Heiss & Chiara Bucaria (eds.), *Between text and image: Updating research in screen translation*, 79–99. Amsterdam/Philadelphia: Benjamins.

Pavesi, Maria. 2009. Dubbing English into Italian: A closer look at the translation of spoken language. In Jorge Díaz-Cíntas (ed.), *New trends in audiovisual translation*, 197–209. Bristol: Multilingual Matters.

Quaglio, Paulo. 2008. Television dialogue and natural conversation: Linguistic similarities and functional differences. In Annelie Ädel & Randi Reppen (eds.), *Corpora and discourse: The challenges of different settings*, 198–210. Amsterdam/Philadelphia: Benjamins.

Quaglio, Paulo. 2009a. *Television dialogue: The sitcom* Friends *vs. natural conversation*. Amsterdam/Philadelphia: Benjamins.

Quaglio, Paulo. 2009b. Vague language in the situation comedy *Friends* vs. natural conversation. In Maria Freddi & Maria Pavesi (eds.), *Analysing audiovisual dialogue: Linguistic and translational insights*, 75–91. Bologna: Clueb.

Richardson, Kay. 2010. *Television dramatic dialogue: A sociolinguistic study.* Oxford: Oxford University Press.

Romero Fresco, Pablo. 2006. Spanish dubbese: A case of (un)idiomatic *Friends*. *The Journal of Specialized Translation* 6. 134–151. Online at: http://www.jostrans.org/issue06/art_romero_fresco.php (accessed 30 April 2011).

Romero Fresco, Pablo. 2009. Naturalness in the Spanish dubbing language: A case of not-so-close *Friends*. *Meta* 54(1). 49–72.

Scott, Mike. 2004. *Wordsmith Tools, v. 4*. Oxford: Oxford University Computing Unit.

Scott, Mike & Christopher Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education*. Amsterdam/Philadelphia: Benjamins.

Seger, Linda. 2009. *Come scrivere una grande sceneggiatura*. Roma: Dino Audino Editore (It. Transl. of *Making a good script great.* 2nd ed. Hollywood, CA: Samuel French Trade).

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. The search for units of meaning. *Textus* 9(1). 75–106. Reprinted as chap. 2 of *Trust the text: Language, corpus and discourse*, 24–48. London: Routledge.

Sinclair, John. 1998. The lexical item. In Edda Weigand (ed.), *Contrastive lexical semantics*, 1–24. Amsterdam/Philadelphia: Benjamins. Reprinted as chap. 8 of *Trust the text: Language, corpus and discourse*, 131–148. London: Routledge.

Sinclair, John. 2004a. *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, John. 2004b. *How to use corpora in language teaching.* Amsterdam/Philadelphia: Benjamins.

Sinclair, John. in press. *Essential corpus linguistics*. Elena Tognini Bonelli (ed.) London: Routledge.

Taylor, Christopher. 1999. Look who's talking: An analysis of film dialogue as a variety of spoken discourse. In Linda Lombardo, Louann Haarman, John Morley & Christopher Taylor (eds.), *Massed medias: Linguistic tools for interpreting media discourse*, 247–278. Milano: Led.

Taylor, Christopher. 2004. The language of film: Corpora and statistics in the search for authenticity. *Notting Hill* (1998) – A case study. *Miscelanea* 30, 71–86. Zaragoza: Departamento de Filologia Inglesa y Alemanna, Universidad de Zaragoza.

Taylor, Christopher. 2006. *I knew he'd say that! A consideration of the predictability of language use in film.* MuTra 2006 – *Audiovisual Translation Scenarios, EU-High-Level Scientific Conference Series*. Online at: http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Taylor_Christopher.pdf (accessed 30 April 2011).

Taylor, Christopher. 2008. Predictability in film language: Corpus-assisted research. In Carol Taylor-Torsello, Katherine Ackerley & Erik Castello (eds.), *Corpora for university language teachers*, 167–180. Berlin/Bern: Peter Lang.

Van Lancker-Sidtis, Diana & Gail Rallon. 2004. Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification. *Language and Communication* 24(3). 207–240.

Warren, Martin. 2006. *Features of naturalness in conversation*. Amsterdam/Philadelphia: Benjamins.