Online film subtitles as a corpus: An *n*-gram approach

Natalia Levshina

Leipzig University

Abstract

This paper investigates online film subtitles as a separate register of communication from a quantitative perspective. Subtitles from films in English and other languages translated into English are compared with registers of spoken and written communication represented by large corpora of British and American English. A series of quantitative analyses based of *n*-gram frequencies demonstrate that subtitles are not fundamentally different from other registers of English and that they represent a close approximation of British and American informal conversations. However, it is shown that the subtitles are different from the conversations with regard to several functional characteristics, which are typical of the language of scripted dialogues in films and TV series in general. Namely, the language of subtitles is more emotional and dynamic, but less spontaneous, vague and narrative than that of normally occurring conversations. The paper also compares subtitles in original English and subtitles translated from other languages and detects variation that can be explained by differences in communicative styles.

Keywords: *n*-grams, register, subtitles, cluster analysis, correlation, odds ratio, deviation of proportions

1. Aims of the study

Online film subtitles are an attractive source of data for corpus linguistics. They are freely downloadable from numerous online repositories for many different languages. Probably the most attractive feature of film subtitles in comparison with other multilingual parallel corpora

(e.g. the Bible translations, proceedings of the European Parliament, the European Union law and the United Nations documents) is that film subtitles are, as a rule, stylistically much closer to informal spoken dialogues than any of these corpora. Fictional film language (original and translated) is characterized by 'prefabricated orality' (Baños-Piñero & Chaume, 2009). This means that screenplay writers try to create film dialogues such that viewers would recognize them as true-to-life speech. Even a brief inspection of film subtitles suffices to see that subtitlers make an effort to achieve this realism, too. As far as lexicon is concerned, film subtitles usually contain few, if any, special terms and archaisms. From a grammatical perspective, subtitles normally contain short simple sentences, questions, exclamations, commands and other features of involved informal communication. As an illustration, consider a fragment from subtitles of the film *The Black Swan* (2010) in (1). This is an example of the SubRip format, which contains the text and information about the time (up to milliseconds) when a caption should appear on and disappear from the screen.

(1) 268

00:33:22,546 --> 00:33:24,109

- Here, hold this. - Yeah, sure.

269

00:33:25,548 --> 00:33:29,219

You must be so excited.

270

00:33:31,080 --> 00:33:32,668

Are you freaking out?

271

00:33:32,703 --> 00:33:33,740

- Yeah... - Yeah?

272

00:33:35,981 --> 00:33:36,814

Oh, it's okay.

Online film subtitles are particularly convenient for the purposes of language comparison because one can easily find subtitles of the same film in many different languages and create a parallel corpus, using the timing information for alignment. Online film subtitles may be preferable to translations of fiction, such as *The Little Prince* by Antoine de Saint-Exupéry or the books about Harry Potter by J.K. Rowling, because one can choose from a wide selection of films of different genres.

There also exist several ready-made collections of film subtitles that are available for download and/or online queries. Perhaps the earliest and largest collection of film subtitles in different languages can be found in the Opus corpus (Tiedemann, 2008).[1] A subcorpus of film subtitles with Czech as the pivot language is also included in InterCorp.[2] A collection of subtitles that are simultaneously available in many diverse languages is being developed by the author for typology and areal linguistics.[3] As for monolingual corpora, English film subtitles represent a part of the New Model Corpus from the Sketch Engine corpus family.[4] One should also mention transcripts of serial television dramas that constitute Mark Davies' Corpus of American Soap Operas.[5]

Monolingual and parallel corpora based on film subtitles have already been used in linguistic research, although the number of studies is still relatively small. Some pioneering work based on film subtitles in ten and more languages has been done by Author (Author, 2015; Author, In press. a; Author, In press b). In psycholinguistic studies, film subtitles in the original language have been shown to be a reliable source of lexical norms, sometimes outperforming other sources (Keuleers *et al.*, 2010).

Despite these attractive features and promising first results, one should be aware of several caveats in using film subtitles for theoretical and applied linguistic research. A well-known issue is the possible influence of the source language on the target language in translations (so-called 'translationese', see Johansson & Hofland, 1994), although this problem is shared by all parallel corpora. In online subtitles, the problem is exacerbated by the fact that it is often impossible to tell which language was the source. For example, a French film can be translated into German directly or via English.

Second, professional subtitlers have rigid rules to follow with regard to the maximum length of a line, the time during which a caption should stay on screen, etc. (e.g. Díaz Cintas & Remael, 2014: Ch. 4; Deckert, 2013: App. 1). Many elements are omitted or reformulated with a preference for shorter constructions. In some films, the percentage of omitted elements can be quite high. For example, speech reduction in Spanish subtitles of one of Woody

[1] http://opus.lingfil.uu.se/ (last accessed 21.04.2015).
[2] www.korpus.cz/intercorp (last accessed 21.04.2015).
[3] [The name of the corpus and URL are omitted for purposes of anonymization]
[4] https://www.sketchengine.co.uk/new-model-corpus/ (last accessed 08.09.2015).
[5] http://corpus.byu.edu/soap/ (last accessed 08.09.2015).

Allen's films was as high as 40% because the dialogue was too fast and verbose to be represented fully on the screen (Díaz Cintas & Remael, 2014: 199). As for examples of reformulation, subtitlers tend to use *will* instead of *be going to*, replace light verb constructions with simple verbs (e.g. *feel* instead of *have the feeling*), and use simple rather than complex tenses (e.g. simple past instead of past perfect) (Díaz Cintas & Remael, 2014: 202–204).

Moreover, previous research into transcribed TV series dialogues and films has revealed a few differences between scripted film or TV series dialogue and spontaneous conversations. In particular, it has been observed that narrative and 'vague' elements and some discourse markers are underrepresented in film and TV dialogues in comparison with naturally occurring conversations (e.g. Mittmann, 2006; Quaglio, 2008; Bednarek, 2011). At the same time, most researchers seem to stress similarity between film or TV series dialogue and naturally occurring conversations, as far as their pragmatic and lexico-grammatical features are concerned (see Dose, 2014: Ch. 4.3.4 for an overview).

Finally, no one can guarantee the quality of film subtitles downloaded from online repositories. Both non-translated and translated subtitles may contain typos and errors. It is very difficult to find reliable information about the subtitler of a specific film and his/her linguistic background and expertise. Although most repositories have systems of downvoting bad subtitles, this does not guarantee a high quality of subtitles that have not been marked as poor. However, in my personal experience based on linguistic analysis of subtitles and native speakers' evaluation of text samples in several languages, the vast majority of subtitles have acceptable quality, although one can see occasional transcription and translation errors (cf. Bednarek, 2010: 70), as well as orthographic and punctuation mistakes. As can be seen from online comments or additional information in the files, many subtitles have been corrected several times, which implies a high level of dedication of the members of the online subtitling community.

Do all these peculiarities make film subtitles too risky to be used for linguistic research? To answer this question, the present paper assumes a quantitative approach, treating English subtitles on a par with other registers of written and spoken English. The main research questions of this paper are as follows:

1) Do English film subtitles represent a language variety that is fundamentally different from other varieties of spoken and written English?

2) How similar are film subtitles to normally occurring conversations?

3) What are the distinctive linguistic features of subtitles in comparison with naturally occurring informal conversations?

4) How similar are subtitles in original English and subtitles translated from other languages? What are the linguistic differences between these types of subtitles?

These questions will be answered with the help of an *n*-gram approach. To answer Questions 1 and 2, I will perform a correlation analysis and hierarchical clustering of registers

based on *n*-gram frequencies. To answer Question 3 about the distinctive linguistic features of subtitles in comparison with naturally occurring conversations, I will analyse the *n*-grams that are distinctive of subtitles and those that more frequently occur in conversations. A similar procedure will be applied in order to answer Question 4. For that purpose, *n*-grams that more frequently occur in original English subtitles will be compared with those that can be found in subtitles translated from French and several other languages.

The remaining part of the paper is organized as follows. Section 2 describes the data and methods. Section 3 presents the results of the clustering models and correlation analyses. Section 4 discusses the distinctive *n*-grams of subtitles in comparison with the spontaneous conversations in British and American English, whereas Section 5 contrasts the original English subtitles with those translated from other languages. Finally, Section 6 summarizes the results.

2. Data

2.1. Corpora

The subtitles investigated in this study were collected from the online repository OpenSubtitles.org.[6] The films represent various fictional genres, according to the genre classification from the International Movie Database[7]: drama, adventure, fantasy, comedy, mystery, crime, etc. In total, I selected 23 files for films that were originally in English, 20 files for films that were originally in French and 14 files for films originally in languages other than English or French (14 different languages). According to the International Movie Database, most English films in the sample were either American or produced by international teams with American participation, e.g. USA/UK/Australia. French was chosen because it is easy to find sufficient data, due to the popularity of French films. It was considered separately from the other languages because one of the goals was to try to identify linguistic features that would reflect the properties of one specific source language (see Section 5). The overwhelming majority of the films, English and non-English, were created in the last 20 years. All files were in the SubRip (.srt) format.

The subtitles were compared with samples from well-known corpora of written and spoken British and American English. Both British and American data were used because subtitles do not represent only one of these two varieties exclusively. In fact, most files in the sample contain the American spelling variants (e.g. *color*), but there are a few files where the British variants are found (e.g. *colour*). Each national variety was represented by two written

[6] http://opensubtitles.org (last accessed 21.04.2015).
[7] http://www.imdb.com/ (last accessed 21.04.2015).

registers (newspaper texts and fiction) and two spoken registers (transcripts of informal conversations and radio and TV broadcasts, which mostly represented unscripted conversations). The choice of registers was motivated by their availability in large comparable corpora of British and American English. The British data were taken from the British National Corpus (BNC). The files were selected on the basis of the meta-information. All files included from 5,000 to 15,000 tokens, which made them comparable to the subtitle files. The American data from newspapers, fiction and media broadcasts were taken from the corresponding components of the Corpus of Contemporary American English (COCA). For each of the components, a local copy of the corpus contained 23 very large files that represent years from 1990 to 2012. A sample of 8,000 words was drawn from each of the 23 files for each register.[8] The informal conversations in American English were taken from the Santa Barbara Corpus of Spoken American English (SBCSAE). The richly annotated dialogue scripts were stripped from the information about pauses, background noises, coughing, etc. with the help of a Python script. The number of files and the number of tokens in each subcorpus are shown in Table 1.

Subcorpus	**Number of files/samples**	**Number of tokens**
Subtitles of films originally in English	23	254,914
Subtitles of films originally in French	20	132,159
Subtitles of films originally in other languages	14	130,552
BrE informal conversations (BNC)	29	268,370
AmE informal conversations (SCBSAE)	19	87,481
BrE radio and TV broadcasts (BNC)	27	234,399
AmE radio and TV broadcasts (COCA)	23	184,000
BrE newspapers (BNC, only national)	24	237,080
AmE newspapers (COCA)	23	184,000
BrE fiction (BNC)	23	192,528
AmE fiction (COCA)	23	184,000
Total	248	2,089,483

Table 1. The number of files/samples and the number of tokens in subcorpora.

2.2. Extraction of *n*-grams

[8] Although the fiction subcorpus of COCA includes film scripts, the sample drawn for this study did not contain them.

The next step was to extract *n*-grams with *n* of 1, 2 and 3 with the help of Python scripts. It did not make much practical sense to try larger *n* because of the relatively small sizes of the subcorpora, which would produce too many *hapax legomena*. Moreover, as Gries et al. (2011) demonstrate, the longer *n*-grams do not add much new information for the purposes of register classification. *N*-grams were defined as sequences of *n* tokens within a sentence, which were not separated by punctuation marks. Punctuation marks themselves were not considered to be elements of *n*-grams. The difference between low and upper case was disregarded. The contracted forms (e.g. *I'll*, *don't*) were regarded as combinations of two grams (*I + 'll* and *do + n't*). The possessive marker *'s* was treated as a separate gram. This method of tokenization was chosen because it had been used in the majority of the selected corpora. In the corpora where this was not the case, the data were first normalized automatically with the help of a Python script.

These *n*-grams and their frequencies in each register served as an input for all subsequent analyses, which are described in Sections 3–5. Since the results based on the 2-grams were intermediate between the ones based on the 1-grams and 3-grams, the 2-grams will not be discussed for reasons of space.

3. Clustering models of English registers

3.1. Methodology

This section represents a series of correlation and cluster analyses based on the frequencies of the 1-grams and 3-grams, with a focus on the relationships between the subtitles and the other registers. The idea of using *n*-grams for comparison of language varieties is not new (e.g. Biber et al., 1999: Ch. 13; Xiao & McEnery, 2005; Gries et al., 2011). This approach was chosen because it is much less labour-intensive than the traditional multidimensional analysis (e.g. Biber, 1988). It is also more objective, as it does not require an *a priori* selection of linguistic features.

The first step was to compute the correlation coefficients. The Pearson correlation coefficient *r* was used because it was shown to be useful for discrimination between the registers with the help of *n*-grams (Gries et al., 2011). In addition, my own experiments with the data have demonstrated that another popular correlation coefficient, the rank-based Spearman *rho*, yields a much less interpretable picture of register variation in English. The Pearson *r* ranges from –1 to 1, where –1 indicates a perfect negative, or inverse correlation,

and 1 stands for a perfect positive correlation. 0 indicates a lack of any relationship. All correlation coefficients that were computed were greater than 0.

The next step was to transform the correlation coefficients into distances by subtracting the former from 1. After that, a hierarchical clustering analysis was performed on the basis of the Ward algorithm, which usually produces compact clusters. A series of hierarchical agglomerative clustering models was created, for 1% of the most frequent grams, 5%, 10%, 25%, all data without hapax legomena and all data with hapax legomena. All these clustering models were nearly identical.

All statistical analyses and graphics presented in this paper were performed or created in R, a free software environment for statistical computing and graphics (R Core Team, 2014). The following subsections 3.2 and 3.3 present the clustering solutions based on the 1-grams and 3-grams.

3.2. A clustering model based on 1-grams

Figure 1 displays a hierarchical clustering model based on all 56,619 1-grams. The figure should be interpreted as follows. The tree 'grows' from the 'leaves' (i.e. the registers) to the 'root' (the top merge). Pairs of leaves or 'branches' (i.e. clusters with several leaves) merge from bottom to top until all leaves and branches are included in the tree. The smaller the distance between two leaves or branches, the sooner they will merge. Therefore, one can expect the registers with similar frequencies of the same *n*-grams to cluster together, and the registers with different frequencies to belong to different clusters.

One can see that the registers are subdivided into two large clusters, one representing the written British and American registers and the TV and radio broadcasts, and the other containing the British and American informal conversations and the subtitles. This can be interpreted as the distinction between more formal and less formal registers. The translated subtitles form a small cluster separate from the original subtitles, although all types of subtitles merge very soon, which indicates a high level of similarity between the original and translated subtitles.

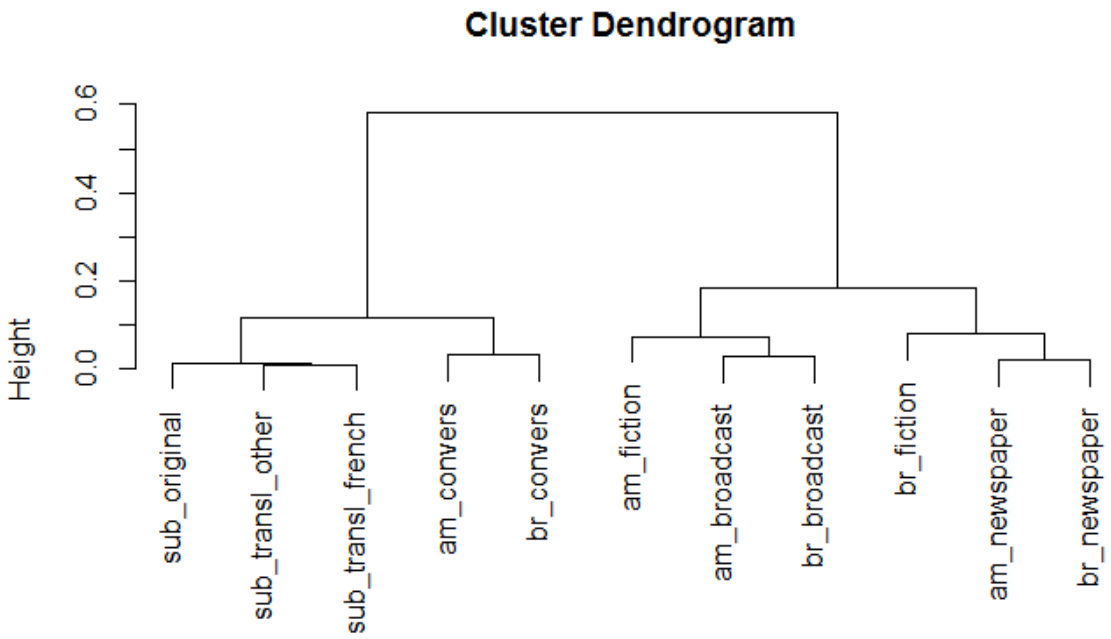

Figure 1. A clustering model based on all 1-grams.

A closer look at the correlation coefficients (see Table 2) reveals additional information. The correlation coefficients between the three types of subtitles are extremely high; they are, in fact, the highest coefficients among all types of registers. This means that there is no principled difference between the subtitles in original English and translations, as far as the frequencies of the 1-grams are concerned. As one could expect on the basis of the clustering model, the strongest correlations between the subtitles and the other registers are observed in the case of the British informal conversations, followed by the American informal conversations (see the highlighted rows in Table 2). This holds for all three types of subtitles, although the translated subtitles tend to have slightly lower coefficients than the original subtitles. The next highest correlations are with the TV and radio broadcasts, which are followed by the fiction. The lowest correlation coefficients are observed between the subtitles and newspapers. For comparison, the lowest correlation between all registers is found between the British conversations and the British newspapers (r = 0.586). In the American data, the lowest correlation is between the conversations and newspapers, too, although the correlation is higher (r = 0.686). This suggests that the differences between the traditional registers are greater than the difference between the subtitles and the informal conversations.

	sub_original	**sub_transl_other**	**sub_transl_french**
am_broadcast	0.88	0.877	0.866
am_convers	0.927	0.904	0.903

am_fiction	0.821	0.82	0.817
am_newspaper	0.693	0.695	0.691
br_broadcast	0.891	0.883	0.871
be_convers	0.947	0.93	0.936
br_fiction	0.701	0.706	0.713
br_newspaper	0.635	0.637	0.629
sub_original	-	0.99	0.988
sub_transl_other	0.99	-	0.991
sub_transl_french	0.988	0.991	-

Table 2. Correlations between subtitles and other registers based on all 1-grams (Pearson's *r* coefficients).

3.3. Clustering models based on 3-grams

Figure 2 displays a clustering solution based on all 965,909 3-grams. The clustering solution displays a large cluster with the spoken data and a cluster with the written registers, except the British newspapers, which merge the last.

The subtitles cluster with spoken registers: they first merge with the informal conversations and then with the TV and radio broadcasts. This is supported by the correlation coefficients displayed in Table 3. Again, the strongest correlations are between the subtitles and conversations, followed by the broadcasts and fiction. The original and translated subtitles again are more similar to one another than to the other registers, although the translations are again slightly less correlated with the other registers than the subtitles in original English.

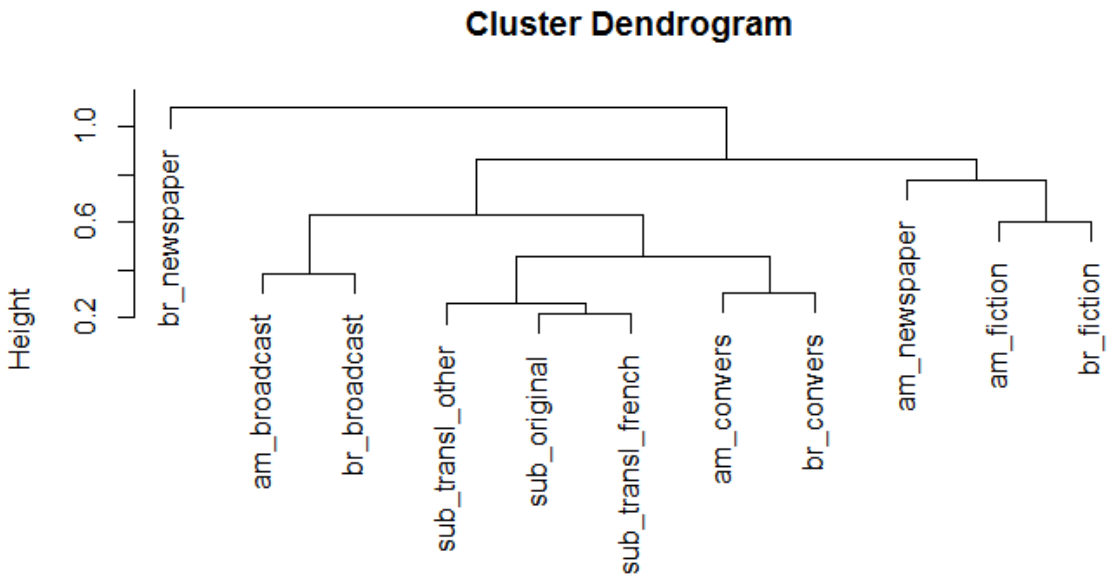

Figure 2. A clustering solution based on all 3-grams.

Register	sub_original	sub_transl_other	sub_transl_french
am_broadcast	0.578	0.505	0.489
am_convers	0.685	0.615	0.599
am_fiction	0.446	0.408	0.4
am_newspaper	0.305	0.26	0.256
br_broadcast	0.542	0.474	0.47
be_convers	0.696	0.651	0.63
br_fiction	0.428	0.402	0.395
br_newspaper	0.099	0.072	0.079
sub_original	-	0.78	0.756
sub_transl_other	0.78	-	0.743
sub_transl_french	0.756	0.743	-

Table 3. Correlations between subtitles and other registers based on all 3-grams (Pearson's *r* coefficients).

3.4. Clustering models and correlations: interim conclusions

Section 3 has discussed several clustering solutions and correlation coefficients that help us answer the first two research questions. The answer to the first question, namely, whether film subtitles represent a variety of English that is fundamentally different from other registers, is

negative. In both clustering models, the subtitles do not form a cluster that would be separate from the other registers. Moreover, the subtitles are more similar to some traditional registers than these traditional registers are similar to one another. From this one can conclude that subtitles represent language that does not differ fundamentally from English produced in more naturalistic settings. In both analyses, the subtitles cluster together with the informal spontaneous conversations. The subtitles also display the highest correlations with this register, in particular with the British informal conversations. These correlation coefficients are in fact the highest among all registers compared, which suggests a close similarity between the subtitles and conversations (cf. Question 2).

In addition, one can make some conclusions regarding the fourth question about the relationships between film subtitles that are translated from other languages and subtitles in original English. The analyses reveal high positive correlations between the subtitles that are translations and those which are not. The corresponding correlation coefficients are in fact the highest observed coefficients between all registers in the data set. This suggests a high level of similarity between different types of subtitles. However, the translated subtitles tend to be slightly less strongly correlated with the other spoken registers than the original ones. Section 5 will discuss possible explanations for these differences.

4. Distinctive *n*-grams in the subtitles and informal conversations

4.1. Methodology: distinctive *n*-gram analysis based on odds ratios and deviations of proportions

This section investigates the third question, which concerns the linguistic differences between the film subtitles and the British and American spontaneous informal conversations. There exist different methods of identifying distinctive elements of subcorpora and registers. One can use multivariate approaches, such as Principal Component Analysis or Factor Analysis, (e.g. Biber, 1988), and identify the factor loadings of different linguistic features on the dimensions of register variation. Another approach is to identify keywords in a subcorpus in comparison with another subcorpus or a large reference corpus (Scott, 1999). Keywords are words that occur in the subcorpus of interest more frequently than one could expect to occur by chance alone. Whether the differences between the observed and expected frequencies are statistically significant is determined with the help of statistical measures, such as the log-likelihood ratio (Dunning, 1993), the *chi*-squared statistic and the Fisher Exact Test *p*-value (see an overview in Baron et al., 2009). Yet another possibility is to compare the rankings of *n*-grams (e.g. Bednarek, 2011) in texts of different registers.

Here I will use an alternative method that compares the relative frequencies of *n*-grams in the subtitles and the British and American spontaneous conversations. This method was

chosen over the traditional multidimensional analysis because this paper focuses on the differences between the subtitles and the spoken data, rather than on the differences between all registers, most of which have been extensively explored. It is also preferable to the ranking comparison approach because the latter involves a loss of information when the level of measurement goes down from the ratio scale to the ordinal scale. The keyword approach based on significance testing has a few problems with underlying statistical assumptions. First, it does not take into account the fact that many *n*-grams are sampled from one and the same text written by a specific author. Therefore, the observations are not sampled randomly. In this situation, the keyword approach based on the computation of a hypothesis testing statistic is problematic. Another problem arises when $n > 1$. Consider a simple example: if a bigram is *in spite*, the chances are high that the following bigram will be *spite of*. The assumption of independence of observations is violated again, but in a different way.

Because of all these problems, I will use a descriptive measure of effect size, rather than a hypothesis testing statistic, for identification of the *n*-grams that are the most distinctive of the subtitles and those that are the most distinctive of the conversations. More specifically, I will use the odds ratio, which is the ratio of the odds of a bigram in one type of text to the odds of a bigram in another type of text. The method is as follows. For every *n*-gram, one needs four scores shown in Table 4:

- *a*: the raw frequency of the *n*-gram in the subtitles;

- *b*: the raw frequency of all other *n*-grams in the subtitles;

- *c*: the raw frequency of the *n*-gram in the informal spontaneous conversations;

- *d*: the raw frequency of all other *n*-grams in the informal spontaneous conversations.

	Frequency in subtitles	**Frequency in conversations**
n-gram	*a*	*c*
all other *n*-grams	*b*	*d*

Table 4. Frequencies required for computation of odds ratios.

The traditional odds ratio is computed according to the formula in (2):

(2) $$\text{OR} = \frac{a/b}{c/d} = \frac{a*d}{b*c}$$

If the odds of an *n*-gram are equal in both registers, the odds ratio will be 1. If the odds of an *n*-gram are higher in the subtitles than in the conversations, the odds ratio will be greater than 1. If the odds of an *n*-gram are higher in the conversations, the odds ratio will be between 0 and 1. The greater the OR, the more distinctive (overrepresented) the *n*-gram in the subtitles, and the less representative it is of the conversations, and vice versa. Since the frequency *c* may equal zero (i.e. when a given *n*-gram does not occur in the conversations), there is a danger of division by zero. To avoid this problem, I will use a 'discounted' version of OR, adding a small number (0.5) to each of the four frequencies.

Another concern is that some frequent *n*-grams may occur in one text only. Such *n*-grams are not representative of the entire register, even if their frequency is very high. Table 5 shows top five 1-grams that occur relatively frequently in the subtitles data, based on their discounted OR. These are proper names of film protagonists. Each of these names occurs only in one subtitle file. Obviously, such information is not particularly informative.

1-gram	**Frequency in subtitles**	**Frequency in conversations**	OR_{disc}
howard	137	0	207.6
malkovich	124	0	188
paro	109	0	165.3
gatsby	95	0	144.2
daisy	93	0	141.1

Table 5. Top five distinctive 1-grams in the subtitles (compared with spontaneous conversations).

To solve this problem, one needs to take into account the dispersion of *n*-grams in the subcorpus. One could filter out the *n*-grams that occur in less than a predetermined number of corpus documents. However, this approach would not take into account the fact that an *n*-gram has higher chances to be detected in a large text than in a small text, based on chance alone. In this paper, I will use an alternative approach, which was suggested in Gries (2008) and Lijffijt and Gries (2012) and which takes into account the differences in the probabilities of an *n*-gram occurrence depending on the size of the corpus components. In this approach one computes a deviation of proportions (DP) score for every word. This measure reflects how much the relative frequencies of a word in different components of a corpus deviate from what one could expect based on the size of each corpus component. The greater the deviation, the more unevenly the word is dispersed and therefore the less representative it is of the corpus as a whole. The formula for the computation of DP is as follows:

(3) $$DP = 0.5 * \sum(\left|P_{obs} - P_{exp}\right|)$$

In this formula, P_{obs} is the proportion of all instances of a word in a given component of a corpus; P_{exp} is the relative size of the corpus represented as a proportion of the number of words in the given component relative to the number of words in the total corpus. P_{exp} takes into consideration the size differences between the corpus components. DP represents the sum of absolute differences between P_{obs} and P_{exp}, divided by 2.

This score is then normalized in order to be distributed from 0 to 1 with the help of the following formula:

(4) $$DP_{norm} = \frac{DP}{1-\min(P_{exp})}$$

If one computes the scores for the words in Table 5, they will range from 0.944 to 0.984. This indicates that the words are dispersed very unevenly. Using an arbitrary cut-off point of DP_{norm} = 0.5, one can be sure that such cases are filtered out and the remaining *n*-grams are truly representative of the subcorpus as a whole. A stricter (lower) cut-off value would be less practical because the number of remaining *n*-grams becomes too small, especially in the case of 3-grams.

In the remaining part of Section 4, I will discuss the results of the *n*-gram analysis for the 1-grams and 3-grams. Since the results of a 2-grams analysis are similar to the ones based on 1-grams and 3-grams, they will not be discussed for reasons of space.

4.2. Subtitles vs. conversations: 1-grams

More frequent in subtitles				**More frequent in conversations**			
1-gram	**Freq. subtitles**	**Freq. conversations**	***OR*$_{disc}$**	**1-gram**	**Freq. subtitles**	**Freq. conversations**	***OR*$_{disc}$**
10	88	0	133.6	*erm*	0	850	<0.001
20	69	0	104.9	*cos*	2	682	0.003
3	125	1	63.2	*er*	6	1343	0.003
promise	94	3	20.4	*eighty*	0	64	0.006
gentlemen	62	2	18.9	*mm*	35	1442	0.018
secret	85	3	18.4	*pound*	6	164	0.03
pleasure	56	2	17.1	*forty*	4	112	0.03
crazy	136	6	15.9	*fifty*	10	136	0.058
sir	411	22	13.8	*quarter*	4	50	0.067
calm	102	6	11.9	*twenty*	33	310	0.081
trust	112	7	11.3	*thirty*	17	158	0.083
act	65	4	11	*tape*	11	92	0.094
protect	49	3	10.7	*round*	29	235	0.094
Mr	593	47	9.4	*twelve*	12	99	0.095
kill	266	21	9.4	*nine*	30	225	0.102

Table 6. Top fifteen most distinctive 1-grams in the subtitles and conversations.

The discussion below is based on the analysis of the top one hundred most distinctive 1-grams in the subtitles with $DP_{norm} < 0.5$ (i.e. the 1-grams with the highest OR) and an equivalent set of 1-grams in the informal spontaneous conversations in British and American English (i.e. the 1-grams with the lowest OR). For illustration, the top fifteen *n*-grams in both registers are shown in Table 6. Note that the prominent positions of numerals (*10, 20, 3*; *eighty, twelve*) in both lists are explained by the fact that numerals are represented in different ways in the two subcorpora: by digits in the subtitles and by words in the conversations. The representation of numbers by digits in the subtitles can be explained by space limitations. The 1-gram *tape* indicates that the conversation participants often referred to the fact of being recorded.

Considering the top one hundred 1-grams that are the most representative of the subtitles, one can observe that this subcorpus contain a relatively high number of direct addresses, attention signals, greetings, and polite formulae, which are exemplified by such 1-grams as *Mr*, *Sir*, *gentlemen*, *kid*, *guys*, *hey*, *thanks*, *excuse*, *welcome*, *sorry* and *pleasure* as in *it's my pleasure* or *with pleasure*. This observation is in line with the one made by Mittmann (2006: 577), who found that TV series dialogues contain more greetings and polite formulae than naturally occurring conversations (see also Freddi, 2012: 392). According to Freddi, film dialogues try to mimic everyday conversations by representing the ritualized acts of daily routine (*Ibid.*). A possible explanation of this difference might be that films and TV drama series represent more dynamic social situations than the informal conversations in the BNC and SBCSAE, where the interlocutors in one recording session usually know one another well and do not come and go often.

Another finding is that the subtitles contain relatively many words that describe a mental state (e.g. *happy*, *sorry*, *scared*, *afraid* and *crazy*), evaluative adjectives (e.g. *beautiful*, *perfect*, *dangerous*, *strange* and *important*) and expletives (*bitch* and *damn*). This corresponds to Quaglio's (2008) observation based on his analysis of TV series *Friends*, where he found the language of the TV series to be more emotional and dramatic than that of normal conversations (see also Bednarek, 2011). This higher degree of emotionality has to do with the entertainment function of films and series. The viewers are supposed to involve and feel with the characters.

Importantly, the subtitles contain a large number of verbs in the base form (e.g. *promise*, *trust*, *act*, *protect*, *kill*, *stop*, *let*, *help*, *speak*, etc.). Most commonly, these verbs are either in the imperative, or in the future tense, or part of an infinitival verbal complement. Consider examples (5) – (7):

(5) *Listen, word to the wise, <u>stop</u> dressing like you're running for Congress.* (*Bad Teacher*)

(6) *I'll <u>help</u> you grab your rocks.* (*Batman and Robin*)

(7) *Let him speak.* (*The Hobbit: An Unexpected Journey*)

Such commands and expressions of intentions create dynamism and propel the plot further.

Looking at the top hundred 1-grams of the conversations, one can find a high number of various discourse markers, such as expressions of solidarity or attention (e.g. *yeah* and *mm*), (dis)fluency markers (e.g. *erm* and *er*), indicators of topic shifts (e.g. *well* and *anyway*) or corrective markers (*actually*). These elements may be less frequent in the subtitles because the latter in fact represent prepared speech, where fewer overlaps, hesitations and corrections can be expected than in spontaneous dialogues (cf. Dose, 2014: 97– 98). Closely related to discourse markers are mental verbs, such as *wonder*, *suppose* and *mean*, which can perform different discursive functions: hedging (*I suppose*), introducing a question (*I wonder*) and clarification (*I mean*). These verbs are also underrepresented in the subtitles. Although one may be inclined to think that discourse markers might be omitted from the subtitles due to space limitations, it has actually been observed that some discourse markers are also underrepresented in transcribed film dialogues, where such limitations are absent (e.g. Mittmann, 2006: 578; Quaglio, 2008: 200). Note that the raw frequencies of the majority of these discourse markers in the subtitles are different from zero. The difference between the subtitles and the conversations is thus only a matter of degree.

In addition to the higher proportion of discourse markers, the conversations have a larger number of past or perfective verb forms (e.g. *had, meant, used, thought, got, walked, bought, stuck, went* and *said*). The spoken subcorpus also contains quite several *ing*-forms (*driving*, *saying*, *putting*, *having* and *sitting*), which often describe the background situation or participants (e.g. *they were having a biology lesson*; *they're like vultures sitting on a rail there*). In addition, the top one hundred distinctive 1-grams include two 3rd person pronouns, *she* and *they*. These features are associated with narrative discourse (Biber 1988). Notably, Bednarek (2011) observes that TV series are also less narrative than normal conversations. This difference can be explained as follows. In film or TV series dialogues, characters usually talk to one another and about their immediate actions and intentions, rather than about past events and third (absent) parties, who may not be immediately accessible to film viewers (cf. Pavesi, 2008: 84–85). Moreover, a story is usually shown as developing in time with the help of visual means, rather than verbally presented by film characters. This conclusion is also supported by a frequent occurrence of time and place adverbials (e.g. *yesterday, then, there, early, week* as in *this week* or *next week*) which are normally used to refer to times and places outside of the current situation (Biber, 1988: 110).

Finally, the subtitles contain relatively few words and constructions that can be described as instances of vague language (Channel, 1994) in comparison with the conversations, where such words and constructions are more frequent. Examples are elements of non-numerical vague quantifiers, such as *(a) bit (of), (a) lot (of)* and *(a) couple (of),* placeholders *stuff* and *ones*, as well as the words *might* and *probably*. Vague language is also underrepresented in TV series in comparison with natural dialogue (e.g. Quaglio, 2008). The speaker can use these elements as hedges, or invite the hearer to construct the meaning together, establishing the atmosphere of informality. Obviously, film language has fewer

vague expressions because of its communicative limitations: the viewers may not always be able to construct the meaning because they are 'overhearers' who have only restricted access to the contextual information 'shared' by the characters on the screen (Dose, 2014: 94–97). Moreover, the viewers do not have an opportunity to ask for clarification if they fail to construct the meaning. In addition, in real conversations speakers may be under time pressure, stress, fatigue, etc. and therefore resort to vague language when they fail to produce an exact expression.

4.3. Distinctive 3-grams

This section presents the results of an analysis of one hundred most distinctive 3-grams in both subcorpora with the normalized DP scores below 0.5. The top fifteen 3-grams are shown in Table 7.

More frequent in subtitles				**More frequent in conversations**			
3-gram	**Freq. subtitles**	**Freq. convers.**	***OR*$_{disc}$**	**3-gram**	**Freq. subtitles**	**Freq. convers.**	***OR*$_{disc}$**
get out of	84	1	39.6	*I du n*	0	82	0.004
I'm here	56	2	15.9	*cos it's*	0	46	0.008
it's me	48	2	13.6	*I said well*	0	43	0.008
let's go	241	12	13.6	*well I'm*	0	43	0.008
out of here	81	7	7.6	*well I do*	0	41	0.008
are you doing	173	20	5.9	*well it's*	1	77	0.014
I was a	46	5	5.9	*oh that's*	1	76	0.014
this is my	46	5	5.9	*and I said*	2	121	0.014
we have a	53	6	5.8	*I mean I*	2	64	0.027
I'm sorry	216	27	5.5	*it's alright*	1	38	0.027
I love you	99	13	5.2	*well that's*	3	83	0.029
take care of	66	9	4.9	*no it's*	2	53	0.033
what kind of	51	7	4.8	*haven't got*	3	72	0.034
where are you	76	11	4.7	*oh it's*	2	48	0.037
I'm afraid	47	7	4.4	*no I do*	1	28	0.037

Table 7. Top fifteen most distinctive 3-grams in the subtitles and conversations.

The analysis of the top one hundred most distinctive 3-grams in the subtitle subcorpus has yielded a few interesting peculiarities. First, the subtitles contain many questions or their elements (e.g. *what is it*, *are you sure*, *what do you*, *why don't*, *how did you*, *are you doing*,

etc.), which mostly express the speaker's reaction to the hearer's actions and help build the conflict situations (Freddi, 2012: 391). There are also very many expressions that contain the verbs of necessity or desire (e.g. *I want you, I need to* and *I'd like*) with infinitival complements (e.g. *I wanted to talk to you*). Both features have been also observed by Bednarek (2011) when she compared dialogues in TV series with other registers. These expressions propel the action forward. They also reveal the characters' motives and feelings and thus make the film viewers identify and feel with the characters.

As for the conversations, one can pinpoint the following peculiarities. First, similar to what has been observed in the case of 1-grams, speakers in spontaneous conversations abundantly use various discourse markers which include (dis)fluency and clarification markers (e.g. *I said well* and *I mean I*), hedges (e.g. *I think they*) and other expressions (e.g. *oh it's* and *tell you what*). The softening and involving functions are also evident in tag questions (*isn't it* and *aren't they*) and downtoners *only* and *just* (e.g. *it's only* and *it's just*).

Again, many distinctive 3-grams in the conversations contain verb forms and adverbials that refer to past events, e.g. *he didn't, and I was, and then I*. There are also elements that introduce reported speech (e.g. *and she said* and *and I said*). These elements are associated with narrative function, which was discussed in the previous subsection. One also finds here a few instances of vague language (*a couple of, a lot of, a little bit, a bit of, it's like* and *something like that*).

4.4. Interim conclusions

The analyses based on 1-grams and 3-grams converge and complement each other. In comparison with the spontaneous conversations, the subtitles contain many expressions that express the speaker's cognitive reactions, desires and emotions. These elements make the story more dramatic, involving and propel the plot forward. The subtitles also contain relatively many greetings, terms of direct address and polite formulae, mainly because the recorded conversations are more static in terms of the communication settings and participants. At the same time, the subtitles have relatively low frequencies of vague expressions, narrative elements and various discourse markers. As for vague expressions, a possible explanation might be that the film audience has only limited knowledge of the context and cannot ask if something is unclear. The relative scarcity of narrative elements in the subtitles may be due to the fact that films usually tell a story by showing it developing on the screen, rather than through someone's monologue. Film characters usually discuss their immediate situations and talk to one another, rather than discuss third parties and past events. Finally, the subtitles also contain fewer discourse markers than the conversations. This can be explained by the lack of actual time pressure in the interaction between the characters, who reproduce prepared text.

Notably, all these features are, in fact, shared by the subtitles with film and TV series transcripts, which were studied by Mittmann (2006), Quaglio (2008), Bednarek (2011), Freddi (2012) and others. Thus, film subtitles represent a type of filmese/serialese.

5. Zooming in on the subtitles

5.1. Introduction

This section compares the original English subtitles with the English subtitles translated from French and then with the translations from the other languages. The *n*-gram approach, which was introduced in Section 4, will be employed to pinpoint the differences between the types of subtitles. If the translations are strongly influenced by the source language(s), one can expect this to be reflected at the level of the top distinctive *n*-grams.

5.2. Distinctive 1-grams

First, I will discuss the most distinctive 1-grams in the original English subtitles and those in the subtitles of French films translated into English. As in Section 4, the analyses are based on a hundred most distinctive 1-grams in each subcorpus with the normalized DP scores below 0.5. The top fifteen 1-grams in each subcorpus are shown in Table 8.

More frequent in original English subtitles				**More frequent in subtitles translated from French**			
1-gram	**Freq. original**	**Freq. transl.**	***OR*$_{disc}$**	**1-gram**	**Freq. original**	**Freq. transl.**	***OR*$_{disc}$**
wondering	19	0	21.8	*Paris*	9	42	0.13
uh	214	6	18.5	*hiding*	5	15	0.199
wow	50	2	11.3	*several*	5	13	0.228
Jesus	47	2	10.6	*arrived*	10	23	0.25
honey	81	5	8.3	*normal*	11	24	0.263
entire	34	2	7.7	*months*	29	57	0.287
sitting	20	1	7.7	*boss*	28	50	0.316
appreciate	19	1	7.3	*hurry*	17	30	0.321
actually	68	5	7	*yesterday*	13	22	0.336
hoping	15	1	5.8	*calm*	29	47	0.348
seriously	25	2	5.7	*hours*	24	38	0.356
oh	799	82	5.4	*broken*	10	16	0.356
wanna	182	21	4.8	*dog*	19	30	0.358

hey	436	51	4.8	*hour*	30	47	0.359
begin	20	2	4.6	*empty*	11	17	0.368

Table 8. Top fifteen most distinctive 1-grams in the original English subtitles (left) and the ones translated from French (right).

An analysis of the top one hundred 1-grams in the original English subtitles shows that this subcorpus contains less formal language than the translated subtitles. Examples are colloquial contractions (*wanna*, *gotta* and *gonna*) and informal exclamations, such as *wow*, *Jesus* and *yeah*. The original subtitles also contain a relatively larger number of discourse markers (*hmm*, *oh*, *uh*, *actually*, *well*, *okay*), as well as polite formulae, greetings, attention signals and terms of address (*thank*, *pleasure* as in *it's my pleasure* or *with pleasure*, *welcome*, *hi*, *hey*, *Mr* and *honey*) in comparison with the translated subtitles. Interestingly, among the most distinctive 1-grams are also a few *ing*-forms (*wondering*, *sitting*, *hoping*, *living*, *putting* and *talking*). One can also find a few instances of vague expressions (*sort*, *thing*, *guess*, *suppose*, *kind*, *lot*, *probably*, *sounds*, *seem*, *might*). The language of the original subtitles is thus more interactive, informal and vague than that of the subtitles translated from French.

The distinctive 1-grams in the translated French films, in contrast, include several past or perfect verb forms (*arrived*, *saw*, *stopped*, *changed*, *sent*, *asked* and *kept*) and 3rd person singular pronouns and verb forms (*he*, *she*, *him*, *his*, *wants*, *needs* and *thinks*). This finding suggests that the language of the translated subtitles is somewhat more narrative than that of the original subtitles. The list also includes the contracted future marker *'ll*. Its higher frequency in the translated subtitles may be explained by the preference of the informal marker *gonna* in the original subtitles.

A corresponding comparison between the original English subtitles and the subtitles translated from other languages (except French) has revealed a very similar picture. In addition, the list of most distinctive 1-grams in the translated subtitles contains the auxiliary *shall*, which is frequently used as a future marker, and conjunction *although*, which is typically used in writing.

5.2. Distinctive 3-grams

This section discusses the most distinctive 3-grams, which were retrieved by using the same methodology. I will begin by comparing the original English subtitles with those translated from French. The top fifteen 3-grams in each subcorpus are shown in Table 9.

More frequent in original English subtitles	**More frequent in subtitles translated from French**

3-gram	Freq. original	Freq. transl.	OR_{disc}	3-gram	Freq. original	Freq. transl.	OR_{disc}
one of those	19	0	20.7	*you'll get*	3	12	0.149
I guess I	15	0	16.5	*I'm scared*	3	9	0.196
'm gonna take	12	0	13.3	*let me go*	8	21	0.21
to meet you	27	1	9.7	*I'll call*	11	27	0.222
and it's	20	1	7.3	*what's wrong*	12	27	0.241
I hope you	20	1	7.3	*look at him*	4	9	0.251
in the world	37	3	5.7	*it's for*	6	13	0.256
thought you were	15	1	5.5	*I'll go*	10	21	0.259
you and I	15	1	5.5	*what's that*	12	24	0.271
I need you	24	2	5.2	*won't be*	11	20	0.298
just don't	23	2	5	*want to see*	13	23	0.305
's what you	13	1	4.8	*it's your*	11	18	0.33
I can get	13	1	4.8	*if it's*	10	16	0.338
some kind of	13	1	4.8	*take care of*	21	31	0.362
ladies and gentlemen	18	2	3.9	*have to get*	12	16	0.402

Table 9. Top fifteen most distinctive 3-grams in the original English subtitles (left) and the ones translated from French (right).

An inspection of the top one hundred most distinctive 3-grams in the original English subtitles reveals the presence of polite formulae (e.g. *ladies and gentlemen*, *I'm sorry* and *to meet you* as in *Pleased/nice/… to meet you*) and a relatively high frequency of hedges, downtoners and attention-getting signals (e.g. *I guess I*, *I think you*, *I'm just*, *you know what).* There are also several elements of expressions that challenge the addressee and propel the plot forward (e.g. *think about it* and *you talking about* as part of *What are you talking about?*). In addition, one can find several vague expressions (*some kind of*, *one of these* and *a lot of*). Finally, the original subtitles have a relatively high proportion of 3-grams with the informal future marker *gonna* (e.g. *'m gonna take*), whereas the subtitles translated from French more frequently contain the future marker *'ll*, e.g. *you'll get* and *I'll call*.

A corresponding comparison between the original English subtitles and the subtitles translated from the languages other than French yields highly similar results and is omitted due to space limitations.

5.3. Interim conclusions

This section focused on the differences between the original English subtitles and the English subtitles of films originally in French and in other languages. The results of the distinctive *n*-

gram analyses do not provide evidence of strong translationese effects. Rather, the main difference lies in the level of (in)formality and interactivity. The original English subtitles contain more discourse markers of different types than the translated subtitles. Moreover, the English original subtitles contain significantly more instances of *gonna*, *wanna* and *gotta*, as well as other informal expressions, while the translators from French and from other languages seem to prefer the more formal form *'ll* (or even *shall*). The vague language is used somewhat more frequently, too, in the original English subtitles, whereas narrative discourse elements are somewhat more frequently used in the non-original subtitles. It seems that the differences between the *n*-grams reflect genre-related or even cultural differences between the countries. It should be mentioned, however, that the differences observed in these analyses are overall more subtle than those between the subtitles in general and the spontaneous conversations (cf. Section 4), as one can conclude from a relatively small number of the corresponding distinctive *n*-grams. The majority of the *n*-grams in the top one hundred lists are lexical units that seem to reflect the plot.

6. Conclusions

This study has compared online film subtitles with other registers of spoken and written British and American English with the help of *n*-grams with *n* from 1 to 3. Different statistical techniques and statistics (hierarchical cluster analysis, correlation coefficients, odds ratios and deviations of proportions as a dispersion measure) were employed. The results of the study are summarized in a concise form below.

1) As the cluster analyses based on the frequencies of *n*-grams have demonstrated, film subtitles are not fundamentally different from other varieties of spoken and written British and American English. The subtitles do not form a separate cluster and merge early with the other varieties.

2) As suggested by the results of the clustering and correlation coefficients based on the frequencies of *n*-grams, film subtitles are very similar to British and American informal spontaneous conversations.

3) In comparison with the informal spontaneous conversations, the film subtitles exhibit a number of differences. First, they contain many emotional expressions (including expletives), and constructions expressing intentions, necessity and desire, which make the viewers involve and feel with the characters and also propel the plot forward. The higher frequency of greetings, polite formulae and direct addresses can be explained by more dynamic social interaction in films than in the recorded conversations. At the same time, film subtitles contain fewer pause fillers, reformulations and other discourse markers, which are typical of spontaneous discourse produced under real-time constraints. Whether and to what extent the creators of film subtitles can further reduce the number of discourse markers for purposes of

compactness, as pointed out by Díaz Cintas & Remael (2014: 214–216), requires a separate investigation. The language of subtitles is also less vague and narrative than that of the informal conversations. These features come from the specific characteristics of films as a medium of communication, where clarity and accessibility of referents play an important role, and where the story usually develops in time with the help of visual means, rather than being explicitly told by characters. Notably, these results are strikingly similar to the results of previous analyses of fictional TV series dialogue transcriptions (e.g. Mittmann 2006; Quaglio 2008; Bednarek 2011).

4) As one can judge from the inspection of the most distinctive *n*-grams, the subtitles of films translated from other languages are not fundamentally different from the subtitles of films that were originally in English, as far as the distribution of the *n*-grams is are concerned. Most differences can be explained by the varying degrees of (in)formality and interactivity. In particular, the original subtitles contain more discourse markers, informal expressions and vague language than the subtitles translated from French and other languages. In this regard, the original subtitles are closer to natural dialogue. However, the language of the translated subtitles is somewhat more narrative, although the differences are very subtle.

As mentioned above, the analyses presented in this paper corroborate the results of the previous studies of film and TV language. It has been found that the distinctive linguistic features of film subtitles are strikingly similar to those of fictional TV series dialogues, which were investigated previously on the basis of TV series transcripts. This finding has two implications. First, since the language of fictional films (the present study) and TV series exhibit very similar peculiarities when compared with the language of spontaneous conversations, one can hypothesize that film subtitles and TV series dialogues belong to one broad register of fictional TV/film dialogue (cf. Bednarek, 2011). Second, since most researchers agree that TV dialogues represent naturally occurring conversations quite faithfully, in spite of these differences (see an overview in Dose, 2014: Ch. 4.3.4), one can conclude that film subtitles can be seen as an acceptable approximation of natural dialogue, as well. The important question of how subtitles are different from the actual film dialogue remains for future research. Another question is whether the above-mentioned linguistic characteristics of film subtitles in English can be extrapolated to other languages and whether one can speak about universal filmese.

To conclude, if film dialogue is a reflection of real dialogue, subtitles are a reflection of a reflection. At the same time, they are remarkably close to actual informal language. The results are of high practical significance for contrastive and typological studies of world languages. The latter is strongly underrepresented in the linguistic data currently used in those disciplines. However, due to the peculiarities described above, it would be risky to use subtitles as data for full-fledged conversational and discourse analysis as a replacement of spoken language (cf. Chaume, 2004; Valdeon, 2008) and filmese in general. For this purpose, comparable original corpora produced in natural settings are indispensable. For other purposes, however, there seem to be no reasons to be overly skeptical, in particular, when one’s approach is based on a quantitative analysis of a large corpus of subtitles.

Corpora

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

The Corpus of Contemporary American English (COCA): 450 million words, 1990 – present. 2008 – . By Mark Davies. URL http://corpus.byu.edu/coca/

Santa Barbara corpus of spoken American English, Parts 1-4. By John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. Philadelphia: Linguistic Data Consortium. URL http://www.linguistics.ucsb.edu/research/santa-barbara-corpus

XXXX [left out for anonymization purposes]

References

Baron, A., P. Rayson, and D. Archer. 2009. Word frequency and key word statistics in corpus linguistics. *Anglistik* 20(1), pp 41–67.

Baños-Piñero, R. and F. Chaume. 2009. Prefabricated orality: A challenge in Audiovisual Translation. In *inTRAlinea Special Issue: The Translation of Dialects in Multimedia.* Available online at http://www.intralinea.org/specials/article/1714 (last access 24.09.2015).

Bednarek, M. 2010. *The Language of Fictional Television: Drama and Identity.* London: Continuum.

Bednarek, M. 2011, The language of fictional television: A case study of the 'dramedy' *Gilmore Girls*. *English Text Construction* 4(1), pp 54–84.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Channel, J. 1994. *Vague Language*. Oxford: Oxford University Press.

Chaume, F. 2004. Discourse markers in audiovisual translating. *Meta* 49(4), pp 843–855.

Deckert, M. 2013. *Meaning in Subtitling: Toward a Contrastive Cognitive Semantic Model*. Frankfurt am Main: Peter Lang.

Díaz Cintas, J. and A. Remael, A. 2014. *Audiovisual Translation: Subtitling*. London/New York: Routledge.

Dose, S. 2014. *Describing and Teaching Spoken English: An Educational-Linguistic Study of Scripted Speech*. PhD Dissertation. Giessen: Justus-Liebig-Universität Giessen.

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

Freddi, M. 2012. What AVT can make of corpora: Some findings from the Pavia Corpus of Film Dialogue. In A. Remael, P. Orero, and M. Carroll (eds.), *Audiovisual Translation and media accessibility at the crossroads*, pp 381–407. Amsterdam: Rodopi.

Gries, S. Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), pp 403–437.

Gries, S.Th., J. Newman, and C. Shaoul. 2011. *N*-grams and the clustering of registers. *Empirical Language Research* 5. URL http://ejournals.org.uk/ELR/article/2011/1 (last accessed 24.09.2015).

Johansson, S. and K. Hofland. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (Eds.), *Creating and using English language corpora*, pp 25–37. Amsterdam: Rodopi.

Keuleers, E., M. Brysbaert, and B. New. 2010. SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods* 42, pp 643–650.

Lijffijt, J., & Gries, S. 2012. Correction to "Dispersions and adjusted frequencies in corpora." *International Journal of Corpus Linguistics*, 17(1), pp 147–149.

Mittmann, B. 2006. With a little help from *Friends* (and others): Lexico-pragmatic characteristics of original and dubbed film dialogue. In Ch. Houswitschka, G. Knappe & A. Müller (eds.), *Anglistentag 2005 Bamberg. Proceedings of the Conference of the German Association of University Teachers of English*, pp 573–585. Trier: Wissenschaftlicher Verlag Trier.

Quaglio, P. 2008. Television Dialogue and natural conversation: Linguistic similarities and functional differences. In A. Ädel & R. Reppen (eds.), *Corpora and Discourse: The challenges of different settings*, pp 189–210. Amsterdam: John Benjamins.

Pavesi, M. Spoken language in film dubbing: Target language norms, interference and translational routines. In D. Chiaro, Chr. Heiss & Ch. Bucaria (eds.), *Between Text and Image: Updating research in screen translation*, pp 79–99. Amsterdam: John Benjamins.

R Core Team. 2014. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* Vienna. URL http://www.r-project.org/.

Scott, M. 1997. PC analysis of key words - and key key words. *System* 25(2), pp 233–45.

Tiedemann, J. 2008. Synchronizing Translated Movie Subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)*.

Valdeon, R. A. 2008. Inserts in modern script-writing and their translation into Spanish. In D. Chiaro, Chr. Heiss & Ch. Bucaria (eds.), *Between Text and Image: Updating research in screen translation*, pp 117–132. Amsterdam: John Benjamins.

Xiao, Zh. and A. McEnery. 2005. Two approaches to genre analysis: three genres in modern American English. *Journal of English Linguistics* 33 (1), pp 62–82.