Use of place names in the subtitle corpus of highest-grossing movies of the past 20 years

Paiders, Janis; Plume, Elina

Source / Izvornik: **Journal of the International Symposium of Students of English, Croatian and Italian Studies, 2018, 43 - 60**

Conference paper / Rad u zborniku

Publication status / Verzija rada: **Published version / Objavljena verzija rada (izdavačev PDF)**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:172:475913

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-05-07**

Repository / Repozitorij:

Repository of Faculty of humanities and social sciences

UNIVERSITY OF SPLIT

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

FILOZOFSKI FAKULTET SVEUČILIŠTA U SPLITU

ZBORNIK RADOVA
MEĐUNARODNOG SIMPOZIJA MLADIH ANGLISTA, KROATISTA I TALIJANISTA

STUDENTSKI ZBOR
SVEUČILIŠTA U SPLITU

ZBORNIK RADOVA MEĐUNARODNOG SIMPOZIJA MLADIH ANGLISTA, KROATISTA I TALIJANISTA

JOURNAL OF THE INTERNATIONAL SYMPOSIUM OF STUDENTS OF ENGLISH, CROATIAN AND ITALIAN STUDIES

Izdavač/Publisher
Sveučilište u Splitu, Filozofski fakultet/
University of Split, Faculty of Humanities and Social Sciences,
Poljička cesta 35, 21000 Split

Odgovorna urednica/Chief
Gloria Vickov

Glavni urednik/Editor-in-chief
Gordan Matas

Urednice/Editors
Ana Ćurčić, Andrea Jović

Recenzenti/Reviewers
Eni Buljubašić, Gordana Galić Kakkonen, Antonela Marić, Nikica Mihaljević, Magdalena Nigoević, Jurica Pavičić, Ilonka Peršić, Antonija Primorac, Anita Runjić-Stoilova, Simon John Ryle, Nataša Stojan, Boris Škvorc, Brian Daniel Willems

Naslovna slika/Cover photo
Dorotea Grgatović

Grafički dizajn/Graphic design
Neda Mandić

Lektura i korektura/Language editing and proofreading
Andrea Jović

Tisak/Print
Redak d. o. o.

Naklada/Edition
150 primjeraka/150 copies

Adresa Uredništva
Poljana kraljice Jelene 1, Split
itheom.split@gmail.com

ISBN: 978-953-352-026-1

Zbornik je objavljen prema odluci br. 003-08/18-06/0015 donesenoj na sjednici Fakultetskog vijeća Filozofskog fakulteta u Splitu dana 25. rujna 2018.

SADRŽAJ/CONTENTS

ČLANCI/PAPERS

Janis Paiders, Elina Plume
University of Latvia
paidersjanis@gmail.com, elina.plume@gmail.com

original scientific paper

Use of place names in the subtitle corpus of highest-grossing movies of the past 20 years

This research analyzes the frequency of toponyms mentioned in the subtitles of the highest-grossing movies in the United States in the last 20 years. The research object consists of 11 subtitle corpora each comprised of subtitles from 50 movies with the highest domestic grosses in the United States in their respective movie genre. The goal of this research is to analyze the frequency and variation of toponyms mentioned in the created corpus. Another aspect this research focuses on is the differences in toponym use between movie genres based on their frequency and other characteristics. Overall analysis of the corpus shows that the most often mentioned country was the USA and the most often mentioned city was New York, however, the overall distribution of place names was diverse and covered most of the countries in North America, Europe and Asia. Countries in South America and Africa were mentioned less often. Movies in the western genre mostly mentioned place names in America, but action and drama genres mostly used international place names.

The results between different genres show that Hollywood movies have varying levels of globalization that is reflected in the used toponyms. Those movie genres that have a higher share of foreign (non-US) box office revenue (e.g. action and adventure movies) use more global and diverse toponyms when compared to more domestic movie genres (e.g. westerns).

Key words: corpus analysis, toponyms, movies, subtitles, frequency analysis.

INTRODUCTION

This research paper focuses on analyzing the frequency of place names mentioned in the subtitles of 1997-2016 highest-grossing movies in the United States. The research object consists of a corpus comprised of subtitles from 550 movies (50 from each of the 11 selected genres) with the highest domestic gross in the United States in 2016. The goal of this research is to analyze the frequency and variation of place names mentioned in the created corpus. The research questions are: (1) which place names are mentioned most frequently in the movie subtitle corpus? (2) does the frequency of certain place names depend on the genre of the movie?

To achieve this goal the research paper is organized into four chapters. The first chapter is about place name classification and identification and it will discuss the characteristics of place names and the most common difficulties that arise when analyzing them. The second chapter will focus on previous corpus-based research of subtitles. The third chapter will explain the methodology used in this research for corpus creation, place name list creation and the research procedure. The fourth chapter will be about the results of the analysis of the place name occurrences.

The analysis of the corpus shows that the most often mentioned country was the USA by a large margin (4755 occurrences out of the total 8748), however, the overall distribution of place names was diverse and covered most of the countries in North America, South America, Europe and Asia. Locations in Africa were mentioned less often.

PLACE NAME CLASSIFICATION AND IDENTIFICATION

This chapter will discuss the characteristics of place names and the most common difficulties that arise when analyzing them. It will begin by describing the discourse functions of nouns, then it will apply these functions to proper nouns and, by extension, place names. Afterwards, it will describe the methods of classifying place names and the most problematic aspects of automatic place name recognition.

In any instance of discourse, one of the most important aspects for coherent communication is information that is shared by

the speaker and the hearer. To relay this information the speaker uses a noun as a reference. Dirven and Radden distinguish three types of references (1999: 90):

1. indefinite reference, in which the speaker assumes that the hearer has no knowledge of the situation mentioned (e. g., in the sentence 'I met a man at a bar yesterday' the man and the bar is mentioned for the first time);
2. definite reference, in which some knowledge of the situation is assumed (e. g., in the sentence 'The man didn't like the music there' the man has been mentioned previously);
3. generic reference, when the speaker refers to the class as a whole (e. g., in the sentence 'He told me that musicians aren't as good as they were before' musicians refer to the whole class, rather than some specific musicians).

Dirven and Radden specify that all proper names are always definite, since they possess inherent uniqueness, therefore even those proper names that do not use the definite article should be treated as a definite reference (1999: 100). The same applies to place names, which are a subgroup of proper names that denote a specific location.

Huddleston and Pullum make a distinction between proper nouns and proper names (2002: 516). Proper names are expressions that are commonly accepted as a name for some entity, e. g., 'the United States of America'. In contrast, proper nouns are 'word-level units belonging to the category noun' and serve as heads for proper names, however, some proper names have common nouns as their head, e. g., in 'the United States of America' the head is the common noun 'States' (ibid.). Proper nouns can sometimes be used to refer to the whole proper name, especially in colloquial speech, e. g., 'I'm from the States'. However, distinguishing proper nouns from common nouns is very reliant on the context due to homonymy, for example, the word 'Sandwich' can refer to both the proper noun denoting a city in south-east England and the common noun indicating an item of food.

Leech distinguishes seven types of meaning for semantic analysis (1981: 9). For words of reference two types may apply (ibid. 12) – conceptual meaning (covers the components that are understood by the literal use of a word, e. g., 'boy' – young, male human) and connotative meaning (a connotation that a word carries, e. g., 'boy' – energetic, loud). However, proper names and by extension place

names carry no conceptual meaning and cannot be separated into components, according to Leidner (2007: 71). While etymologically place names might be separated into components (e. g., Birmingham - 'Beormund' + 'ingas' + 'ham', meaning 'the farm of Beormund's people'), the conscious awareness of these components is not something that people have when they hear the place name. Therefore, the place names only carry connotative meaning.

One aspect that should be taken into consideration when determining occurrences of place names in subtitle corpora is that place names are not constant and are liable to change due to political and territorial reasons. One example would be Chemnitz – a city in the Eastern Germany that was renamed to Karl-Marx-Stadt during the German Democratic Republic times, but the original name was later renewed (O'Brien 2003: 342). These changes must be taken into account in the research of subtitle corpora, since movies in the historical genre are prone to the use of historical or time appropriate names (e. g., using 'the Soviet Union' rather than 'Russia').

Another aspect that might create difficulties when determining place names is topological metonymy, wherein a place name is used to refer to a more complex concept. Leidner gives an example that is often used in journalistic style – "Washington said...", which refers to the US government that is located in Washington DC (2007: 72). This creates numerous instances in which the subtitle corpus has to be analyzed in-depth, evaluating the context and the speakers, to distinguish the location from other possible meanings.

To accurately distinguish place names, the process of recognizing spatial language in text documents is used. It is more commonly called geoparsing (other terms include geotagging, georecognition, and toponym recognition) (Leidner & Lieberman 2011: 5). In practice, geoparsing has difficulties because of the many ambiguities present in natural language, including ambiguities related to place names. Many non-locations share names with locations e. g., 'Split' can refer to 'Split, Croatia', but might also refer to the verb 'to split'. Removing geo/non-geo ambiguity is crucial for any successful geoparsing attempt. It must be noted that the geoparsing process for country-level place names is often easier than the one for city-level place names, mostly because of the smaller number of country place names, which provides fewer opportunities for geo/non-geo ambiguity. Geoparsing must also deal with misspellings or errors in

the documents themselves (e. g., subtitle corpus) (Leidner & Lieberman 2011: 5).

The use of place names in movies has often been analyzed using in-depth analysis, for example, analyzing the usage patterns in a single film. A research paper by Baqué (2013: 1) analyzes the representation of America as seen by a European in Stanley's Kubrick's 'Lolita' where the movie can be understood as a fictionalized travelogue. The author analyzes the visual effect used in the movie as well as its impact and the places mentioned. In his research he notes that 'America exists in the mind of the audience as a collection of pre-existing images glimpsed in previous films' (ibid. 4), distinguishing how movies use the connotative meaning of place names to offer something familiar to the viewer.

PREVIOUS CORPUS-BASED RESEARCH OF SUBTITLES

This chapter will discuss previous research of subtitles as a corpus. It will begin by describing research of spoken word corpora, subsequently describing subtitle corpora as a cost-efficient method of analyzing spoken word patterns. Afterwards, it will describe some of the research done on subtitle corpora and the possible legal issues that might arise from using subtitles. Finally, this chapter will describe some recent discoveries about evaluating the word frequency in a corpus that consists of numerous distinct documents.

Research done on spoken word frequencies is much needed because it deals with words that are encountered in everyday life (e. g., words related to eating, clothing, furniture, casual social interactions, etc.). Ideally, such spoken corpora would record everything people listen to and say during their day (New et al. 2007: 662). However, making such a corpus would be too costly (ibid.).

One source of transcribed spoken text freely available on the internet are the subtitles of films and television programs. This type of corpus deals with spoken interactions between people in a visible setting and it also encompasses much of the daily heard spoken language, because for many people films and television programs are a substantial part of their language input, as each day people spend 3-4 hours on average watching television (New et al. 2007: 662).

No previous research on place names in subtitle corpora could be found, as subtitles are mostly used to study word frequencies and

parallel translations. In one research regarding the use of movie subtitles to estimate word frequencies, New et al. obtained more than 9474 subtitles for different movies and television series in French language, creating a corpus of more than 50 million spoken words (2007: 662). The research allowed them to produce a detailed analysis of the daily-heard spoken language, but also had a slight bias towards the spoken word frequency of police-related matters that were frequent in the analyzed subtitles of movies and television shows. The results showed that the current subtitle frequency measure seems to be a useful addition to the existing spoken and written frequencies and this kind of corpus can be obtained without the need for manual transcription (ibid. 676).

Keuleers, Brysbaert & New conducted a similar research by creating a subtitle corpus for Dutch language (2010: 643). The corpus included 8443 subtitles, most of which were translated subtitles from American movies and television series. Their research also showed how easy it is to make a good word frequency list for a language by obtaining and analysing subtitles as a corpus. In addition, their research discussed the legal issues that need to be addressed in this type of research (ibid. 649). When large amounts of subtitles are obtained, it is impossible to determine the origin of each subtitle file. Most subtitles available on the Internet appear to fall into two categories: copies of the original subtitles available on DVD or other media, or translations or transcripts made by interested persons, also called fan-created subtitles or 'fansubs' (ibid.). Because of this, the legality of the acquired subtitles might be called into question.

It must be noted that providing subtitles for downloading without permission may be a violation of copyright laws in several countries. When these subtitles are analyzed for research purposes, it is not a violation of copyright and is considered to be fair use of copyrighted material. However, any such research is also indirectly benefiting from a potentially illegal activity and therefore ethical issues should also be considered (Keuleers, Brysbaert & New 2010: 649).

Limon analysed the quality of subtitle translations and noted that most of the translations available online are done by amateur translators and enthusiast, rather than by professional translators (2012: 189). He also notes that while there exist several websites aimed purely at discovering the mistakes made in translations, the

overall quality of translations was satisfactory (Limon 2012: 199). He mentions that the most frequent criticism of translations is aimed at the direct translations of movie titles, which are mostly direct translations of the source material (ibid.).

New et al. distinguish two ways of analyzing the word frequencies of a corpus (2007: 663). The first is by simply calculating the frequency of all different word forms that are encountered in the corpus. This option is simple to use, but causes some interpretation problems because, for example, the word 'play' can be both a verb and a noun, and by knowing just the frequencies, there is no way to differentiate between the two (ibid.). The second option is by parsing the sentences, which allows the researcher to know which grammatical function each word has. This is called a tagged corpus and currently there are many high quality parsers available (ibid. 664).

Adelman, Brown, & Quesada illustrated one of the most recent developments in word frequency analysis in corpus by the finding that the number of times a word occurs in a corpus is less informative than the number of documents in which the word appears (2006: 2). If a word appears in several different documents, it is a lot more memorable than if it appears in a single document numerous times due to the fact that the separation of time and context creates a more memorable imprint in the reader's mind (Glenberg 1979: 95). This was taken into account when analyzing the results of this research and the number of movies that mention a specific place name was determined alongside the absolute frequency of occurrences in all movies.

METHODOLOGY

In order to achieve the goal of this study, a procedure was developed and followed to create a subtitle corpus of 50 highest-grossing movies of the past 20 years (1997-2016) in 11 different genres, as well as a list of place names that are most likely to occur in the corpus. The corpus and the list were optimized for frequency analysis by removing unnecessary information for a cleaner and more comprehensible output. Afterwards, place name occurrences in the corpus were determined using AntConc software, and the resulting list was exported to Microsoft Excel for detailed analysis.

Corpus creation

The creation of the corpus was done according to the established methodology by Sinclair to ensure a representative and homogenous corpus that would accurately represent a portion of movie subtitles (2005: 1-16).

First, the highest-grossing movies of the past 20 years in each genre were determined using the USA domestic grosses as listed on the website BoxOfficeMojo.com. The genres that were looked at were action, adventure, comedy, crime, horror, western, drama, sci-fi, war, history and musical. The total number of subtitles was 550 (50 in each genre), however, due to the fact that some movies have a genre overlap, the number of distinct movies was 407. An example of the overlap can be seen in table 1, where the most frequently overlapping genres are shown (action-adventure, action-sci-fi, adventure-sci-fi, war-history).

	Action	Adventure	Comedy	Crime	Horror	Western	Drama	Sci-Fi	War	History	Musical
Action	X	31	3	3	0	0	3	30	1	1	0
Adventure		x	12	1	0	0	6	23	0	0	1
Comedy			x	4	0	0	5	4	0	0	6
Crime				x	0	0	5	1	0	0	1
Horror					x	0	1	2	0	0	0
Western						x	2	0	0	0	0
Drama							x	6	4	4	1
Sci-Fi								x	0	0	0
War									x	19	2
History										x	1
Musical											x

Table 1. Highest-grossing movie overlap by genres

The subtitles were downloaded from the website Subscene.com because this website offers the option to filter the subtitles by language and closed captioning. Closed captioning subtitles were preferred for the corpus creation because they include extratextual information, such as a character speaking in a different language. An example of this can be seen in figure 1.

```
1
(BUCKY SCREAMING)

2
(CONTINUES SCREAMING)

3
-(KARPOV SPEAKING RUSSIAN)
-(PANTING)

4
Longing
```

Fig. 1. Example of closed captioning in ‘Captain America: Civil War’ subtitles

Afterwards, the 550 subtitle files were modified to remove the timestamps and unnecessary paragraph breaks to conserve space and remove information that would not be needed during the analysis of concordance. This was done in Microsoft Excel by removing lines that contained the ‘-->’ sequence, which is only used in the subtitle format to indicate the length of time a line is seen on the screen. Example can be seen in figure 2.

```
3
00:01:19,040 --> 00:01:20,679
(BEEPING)

4
00:01:30,520 --> 00:01:31,840
- Mama!
- LYRA: We know.
```

Fig. 2. Example of the subtitle formatting in ‘Rogue One’ subtitles

Unfortunately, opening an .srt format file in Excel converts all commas into the tab character which moves the text following the comma to the next cell. This distorted the concordance and made it impossible to extract the final results to a readable Excel file. Because of this, all tab characters were replaced by a whitespace character using Microsoft Word. Furthermore, an utterance was often preceded by a dash to indicate multiple characters speaking in one subtitle fragment. Since in Excel a cell starting with a dash indicates a

formula, all instances of '-' were preceded by an apostrophe, because it indicates that the cell contains text, rather than formulas. Finally, all .srt format files were saved as tab delimited format files (a subtype of .txt format) for accurate importing into Excel and AntConc. The final corpus was 550 .txt format files that were grouped and analyzed by each genre.

Place name list creation

In order to cover most place names that could occur in the movie subtitle corpus, but still keep the list feasible within the limits of computer processing power, several criteria were chosen:

1. all continents and countries were included;
2. all cities with a population of more than 100,000 were included;
3. all U.S. states were included;
4. all locations that consist of fewer than 4 letters were removed to avoid too many false positives;
5. all languages that are spoken by more than 50 million native speakers were included to determine additional mentions of a location, e. g., a character speaking in a specific language;
6. some often used synonyms were included (e. g., 'United States', 'America').

After following all the criteria, the list of place names contained 3907 items, which was within the computational limits of an average computer.

The list of countries and cities was taken from the 'Demographic Yearbook 2015' made by the United Nations, however, during the corpus analysis several inconsistencies and missing locations were noticed and were added manually when found (e. g., Vietnam and Boston are not included in the UN list). The list of languages spoken by more than 50 million native speakers was taken from World Heritage Encyclopedia article in 2015.

Research procedure

The automated method relied on using AntConc software (Anthony 2016) to create KWIC (key word in context) concordance lines. This was done after data scrubbing to ensure that the material was coherent

and did not contain unnecessary information, which would take up place in the limited space that shows the context around the keyword.

The software found more than 20,000 place name occurrences, which, however, also contained numerous false positives because of ambiguities related to place names, which were later manually removed. The list of place names in context was later exported to Excel where the false positives were manually removed and the locations were grouped by country, further distinguishing the type of place name – continent, country, city or US state. The final list had 8748 items.

Due to the fact that the final list of all place name occurrences and the context in which they occur in the corpus contained more than 637,000 words, it would not have been plausible to do a reliable in-depth analysis of them all. Therefore, most occurrences were analyzed by the frequency in which they occur, while those that were mentioned more frequently or often appeared in a more notable context than just a location in which the action takes place were analyzed separately.

RESULTS

The revenue share of the movies from the dataset in total US box office revenue has been increasing steadily (from 28,0% in 1997 to 51,1% in 2017). This is mainly related to the overall yearly revenue increase of movies represented in the dataset (due to ticket price inflation). Also, the emerging role of blockbuster movies (e. g. big budget comic book movie adaptations) is a noticeable factor.

Most (54%) of the 8748 mentioned place names are related to places located in the US. These results show that movies in the analyzed dataset tend to be heavily US-centric (figure 3). However, this trend differs between the 11 analyzed genres. In drama, western and horror genres, more than 70% of all place names are related to US place names. Meanwhile in war genre only 38% of place name occurrences are related to the US, mostly because of movies related to World War II and, to a lesser extent, recent wars in Afghanistan and Iraq. A significant share of analyzed movies in the war genre (e. g. ‘Saving Private Ryan’, ‘American Sniper’, etc.) are still heavily US-centric in terms of who the main protagonists are (US soldiers), but

the wide geographical distribution of these conflicts allows the main action to happen outside the US.

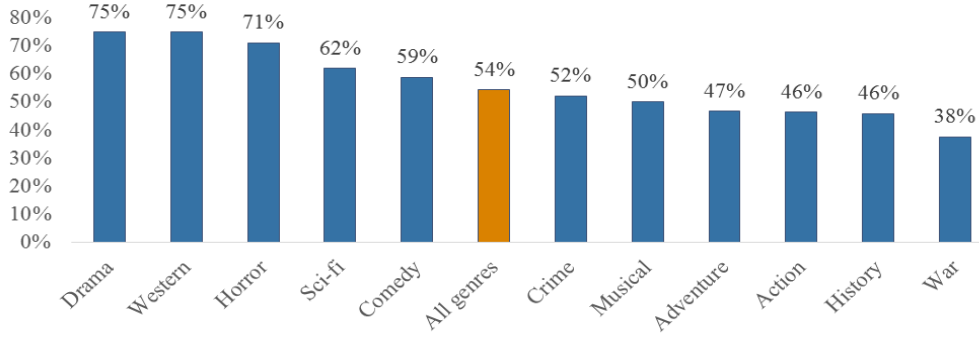
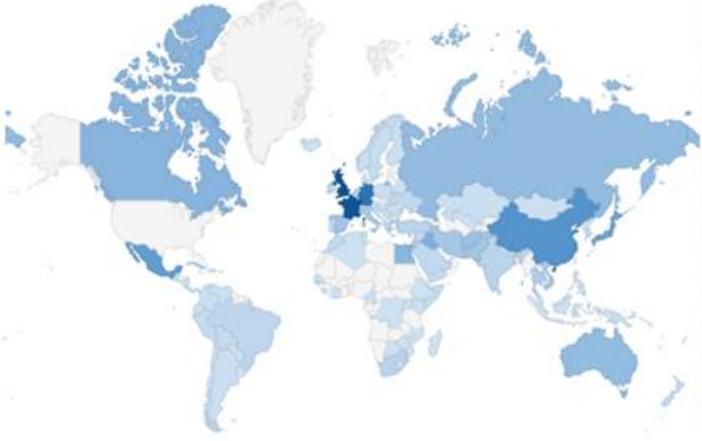


Fig. 3. Percentage of US place names by genre

Countries outside the US are mentioned a lot less frequently in movies of the analyzed data set (figure 4). Place names in the second most often mentioned country (United Kingdom) are mentioned almost 10 times less than those in the US. Even though the difference in occurrences between the US and other countries is this large, the overall number of occurrences is large enough to allow the obtaining of meaningful results about place name mentioning patterns for other frequently mentioned countries.



Country	Occurrences
USA	4755
UK	483
France	470
Germany	317
China	202
Mexico	180
Japan	166
Israel	140
Iraq	132
Canada	119
Total	**8748**

Fig. 4. Frequency of place names in the top 10 most mentioned countries (USA has been removed form visual representation as an outlier)

Germany is one of the countries where occurrences are heavily focused on specific genres. Almost all occurrences are in either war or history genres, or, more specifically, in movies about the World War II (movies in these two genres also have a tendency to overlap). When looking at German place names in the dataset, 40% of all occurrences are the capital city of Berlin and 39% of occurrences are for 'Germany'. Other German cities are mentioned a lot less frequently (only a couple times), e. g. Hamburg, Leipzig, Siegen, Stuttgart, Aachen, etc.

In total, Germany was mentioned in 43 movies (11% of all movies in the dataset) and in the top five movies where Germany gets mentioned most often, three movies are about World War II ('Valkyrie', 'The Monuments Men', 'Red Tails'), one is about the Cold War ('Bridge of Spies') and one is a Marvel blockbuster ('Captain America: Civil War').

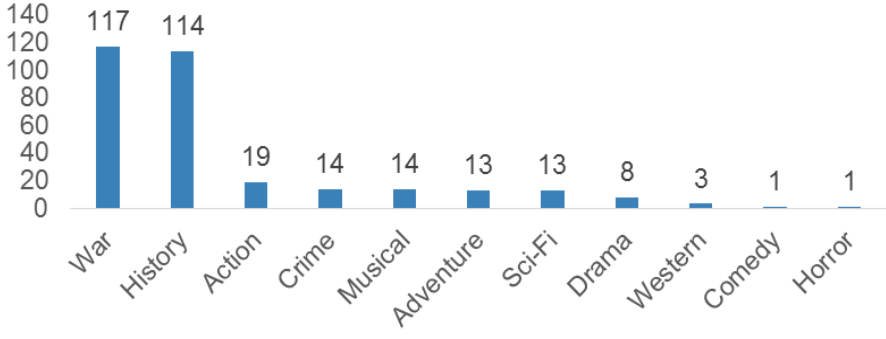


Fig. 5. Frequency of Germany related place names by genre

In contrast to Germany (figure 5) the genre distribution in the occurrences for place names related to France (figure 6) is different. Overall, France is mentioned a bit more often than Germany (figure 4), but its occurrences are scattered across many movie genres. Just

like Germany, France is most often mentioned in history and war movies, but with a lesser emphasis on World War II.

In total France is mentioned in 93 movies (23% of all movies in the dataset) and 58% of all occurrences are for the capital city Paris. The frequency of occurrences for 'France' out of French place names is 36%. That means that only 5% of all place names are for different places in France (e. g. Strasbourg, Caen, Marseille, Bordeaux, etc. that get mentioned a couple of times).

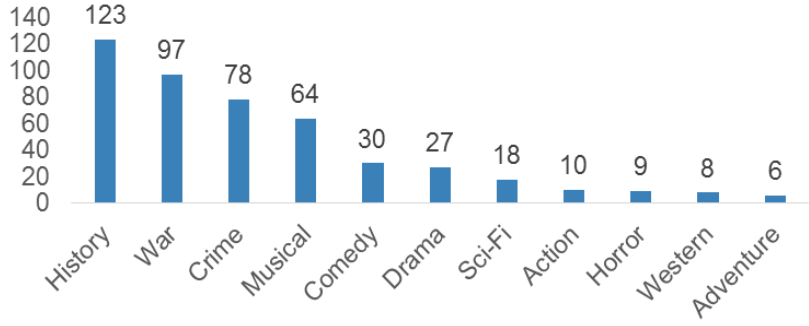


Fig. 6. Frequency of France related place names by genre

Since most of the place names in the analyzed dataset are in the US, their frequency between different US states was also researched in detail. Two of the most US-centric movie genres are westerns and horror movies and they exhibit different characteristics for their geographical distribution of place names.

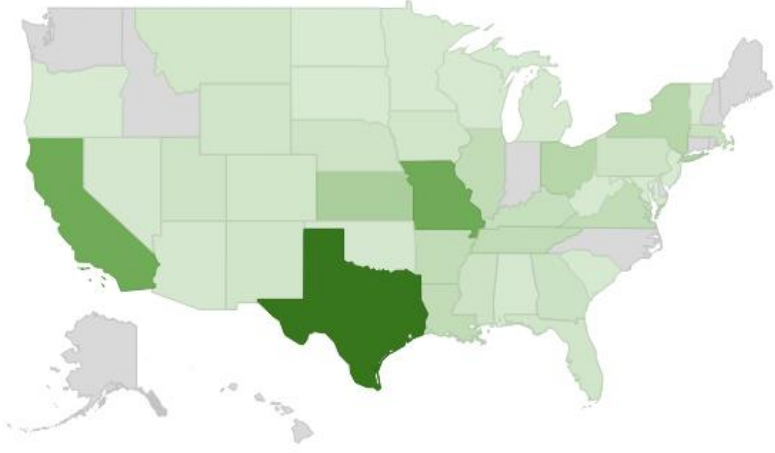


Fig. 7. Frequency of USA related place names by state in the western genre

Western movies are heavily focused on one state in particular, Texas, with 22% US place names in this genre being located in this state (other top states in this genre, California and Missouri, each get mentioned in 10% of US place names in this genre). 26 out of 50 highest grossing westerns mention Texas (figure 7) and movies like 'Texas Rangers', 'The Alamo' and 'True Grit' each mention Texas place names more than 20 times. Out of all occurrences of Texas in the western genre, 120 (76%) of these occurrences are for the state name, while cities like Houston, Lubbock, Austin, Brownsville and Dallas only get mentioned a couple of times.

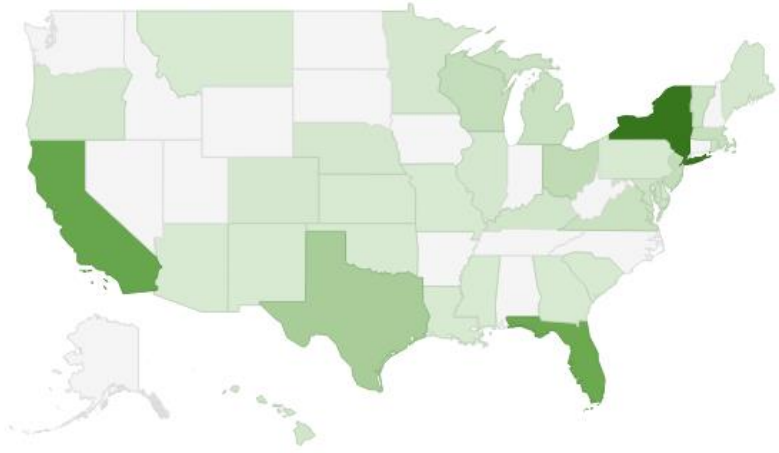


Fig. 8. Frequency of USA related place names by state in the horror genre

Horror genre has a different pattern when compared to westerns (figure 8). The frequency of place names by state is closely related to the population of each state, since the four most populous states (California, Texas, Florida and New York) are mentioned most

often. Horror genre is also specific because even though most of the movies (43 out of 50) mention place names and 35 out of 50 mention ones located in the US, the overall frequency of place names in this genre is low. For US place names, only 7 movies have more than 10 US place name occurrences, showing that the geographical setting of the movie is secondary in this genre.

CONCLUSIONS

There is a significant difference between movie genres in the characteristics of place names used (frequency, mentioned countries, etc.). Some of the disparity between genres can be explained by the role of 'foreign' box office revenue. Genres that are heavily reliant on revenue outside the US (action, adventure, etc.) tend to mention places outside the USA more frequently. The role of non-USA place names in Hollywood movies will probably continue to increase along with the growing role of foreign markets and the decrease of domestic revenue (figure 9).

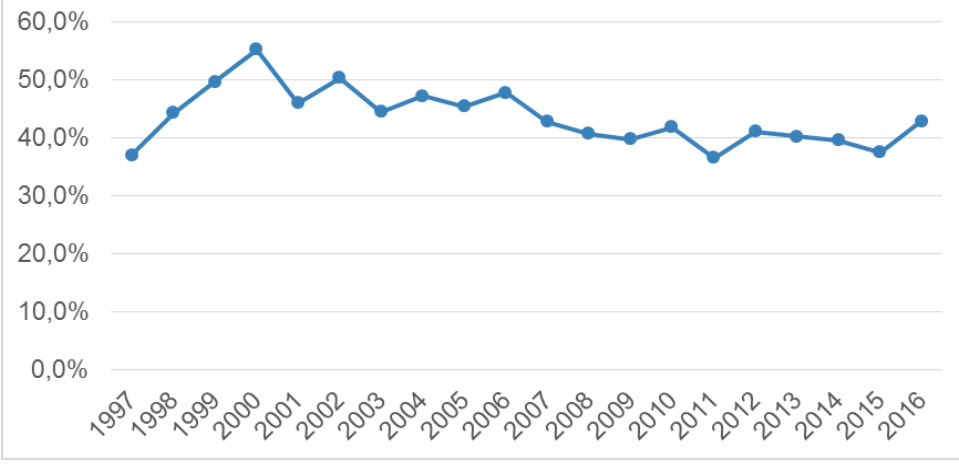


Fig. 9. Changes in the share of USA box office revenue (for the analyzed dataset)

Since 2000, the share of US box office revenue in 50 top-grossing movies has been slowly decreasing. It must be noted that the analyzed movies already have the best box office results in the US. This decrease of US share in revenue is mostly related to the increase of some 'foreign' markets (particularly China in the recent years).

For all 407 analyzed movies 42% of their revenue came from the US, which means that the frequency of US place names (54%) is higher than the share of US revenue. Perhaps this disparity can be explained by different movie-going habits between the US and the rest of the world, where a US-centric movie can still be a financial success in the rest of the world, while US moviegoers are less likely to watch a movie that does not focus on the US. However, in order to test this hypothesis, additional research must be done.

Works cited

Adelman, James S., et al. (2006). 'Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times.' *Psychological Science*, vol. 17, no. 9, 814–823.

Anthony, Laurence (2016). *AntConc (Version 3.4.4).* Tokyo: Waseda University.

Baqué, Zachary (2013). 'Putting the geography of the United States into motion': Kubrick's Lolita as an American Travelogue.' *Miranda*, no. 3.

Dirven, René, and Radden Günter (1999). *Cognitive English Grammar*. Amsterdam: Benjamins.

Glenberg, Arthur M (1979). 'Component-Levels theory of the effects of spacing of repetitions on recall and recognition.' *Memory & Cognition*, vol. 7, no. 2, 95–112.

Huddleston, Rodney D., and Geoffrey K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Keuleers, Emmanuel, et al (2010). 'SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles.' *Behavior Research Methods*, vol. 42, no. 3, 643–650.

Leech, Geoffrey N. (1981). *Semantics: the study of meaning*. Penguin Books.

Leidner, Jochen L. (2007). *Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names*. Dissertation, The University of Edinburgh.

Leidner, Jochen L., and Michael D. Lieberman (2011). 'Detecting geographical references in the form of place names and associated spatial natural language.' *SIGSPATIAL Special*, vol. 3, no. 2, July, 5–11.

Limon, David (2012). 'Film titles and cultural transfer.' *Cultus: the Journal of intercultural mediation and communication*, vol. 5, 189-208.

New, Boris, et al. (2007). 'The use of film subtitles to estimate word frequencies.' *Applied Psycholinguistics*, vol. 28, no. 04, 661–677.

O'Brien, Donough (2003). *Fame by chance: an A-Z of places that became famous (or infamous) by a twist of fate*. Honiton: Bene Factum.

Sinclair, John (2005). 'Corpus and Text - Basic Principles.' In: Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books, 1-16.