Computational Linguistics and Audio-Visual Readability: Analysing Linguistic Features of Intralingual-Subtitles *Corpora*

Andrea Verdino *andrea.verdino@studenti.unimi.it*
Foreign Languages and Literature Department
Università degli Studi di Milano
Milan, 20122, Italy

Abstract

This paper aims to evaluate a new perspective for computational linguistic studies, determining if it is possible to carry out a complete linguistic analysis on tendencies and features of audio-visual works (films, TV-shows, documentaries, videogames) by collecting original intralingual subtitles in POS-tagged linguistic *corpora*. Two corpora were made starting from textual materials of two *'Netflix Original'* TV-series, in the attempt to provide a structured approach for further research on the linguistic aspects of subtitles.

Keywords: Computational Linguistics, Corpus Linguistics, Audio-Visuals, Subtitles, Sketch Engine.

1. INTRODUCTION

The academic relevance of the so-called 'audio-visuals' (movies, series, video games) in linguistic studies is relatively recent and regrettably poor, despite the wide range of researching tools and the vast opportunities that studies based on visual works can offer to investigate the tendencies of contemporary English language. Nevertheless, it is clearly evident that, during the last two decades, the number of viewers of movies and TV-shows (and video gamers consequently) drastically increased. This is due to the fact that *streaming platforms* such as *Netflix* or *Amazon TV* are worldwide available and 'user-friendly', but consideration should be given to the methods and ways in which twenty-first century audio-visuals communicate with their audiences. Nowadays, it is not uncommon to find non-native English speakers watching shows in English (usually helping themselves in the comprehension of unclear passages with valid subtitles) and language, consequently, changed its patterns and structures as a result of contaminations and actualizations.

To carry out the research on the linguistic analysis for the subtitles of the two series, two separate *corpora* were built. A linguistic *corpus* (the Latin word for 'body'), according to Kübler (2005), is a «large collection of texts, especially if complete and self-contained» [1]. Observing the linguistic features and tendencies of a language in a corpus it is possible to extract a series of information about the texts taken into consideration (i.e. word frequencies, lexical preferences, concordances, sociolinguistics features, …). It is common practice considering nearly only written texts worthwhile for a corpus-based analysis, while it is more difficult recording spoken language faithfully. More recently, the majority of scholars and corpus linguists argued with that perspective and consider spoken language as important as the written one for corpus-based studies for specific purposes (Schmidt, Wörner, 2009) [2] while it is true that the amount of spoken-recorded texts for corpora is far less extended than written ones (Newman, 2008) [3].

A different truth can be told about the field of audio-visuals. In fact, starting from the point that everything that is spoken in front of a camera for a general audience (film, series, video game, informative video) it was previously written by someone (either a writer, a screenwriter, or a *scene writer*), the number of written texts that can be collected is nearly endless. Recording studios and dubbing agencies receive the *transcriptions*, or screenplays, and re-write them to compile the

subtitles. As regards the subtitling practice, the general belief is that subtitles are only written for that specific *target language* or TL [4] (the language in which the transcription is translated, they are called *interlingual subtitles*). However, dubbing studios rely on translators and adapters to write at the same time subtitles in the original language of the *source language* or SL. This is partly because transcriptions, while are more interesting for movie-directing scholars and filmmaking students, are not as relevant as subtitles for *AVT* (audio-visual translation) professionals: in fact, transcriptions are not *tagged* (they do not have the right format to be put in the subtitling software) and contain various information that are not relevant (or even incorrect) to write a good subtitle [5], such as the name of the character speaking in that moment and some indications about how he or she speaks and moves in the scene. The second reason behind the importance of writing intralingual subtitles is that every day a larger amount of non-native English speakers prefers watching movies and series in the original language [6] and for most of them it would be difficult to understand every part of speech (considering the differences of accents and speaking speed-ratio of the different characters and hard-comprehension spellings of some words). Some different considerations should be made for the case of subtitles for deaf and hard-hearing people, but they will be omitted for the purpose of the research presented in the paper, whereas they cover a different branch of subtitling studies [7].

Subtitling is not only the pure 'writing of speech'. It is a complete, complex, rule-governed system in which a number of linguistic features needs to be considered for an efficient subtitle. It has been proved that collecting interlingual subtitles in *parallel corpora* (in which the texts are translated in two or more languages, frequently they are *bilingual corpora*) can improve and accelerate a translator job in finding matching strings and concordances for the source and target language, discovering how some particular features that recur in a movie or series were translated or adapted (Bywood et al., 2013) [8]. However, little attention has been given to subtitles corpora that consider only intralingual subtitles, which are the most valuable subtitling method for specific purposes, like language teaching or linguistic analysis (Bird, Williams, 2002) [9]. Demonstrating how researchers, teachers and scholars could benefit from the use of intralingual subtitles corpora can help linguists to meet new perspectives for computational linguistics research.

2. RESEARCH OVERVIEW

Studies on the analysis of audiovisuals through the use of subtitles corpora are recent and almost entirely focused on the construction of parallel corpora, which are a valid instrument for screenwriters and AVT professionals; however, the relevance of intralingual subtitles is confined to a minor role, despite the potentials that intralingual subtitles may have for linguistic analysis.

Lison and Tiedemann (2016) pointed out that «movie and TV subtitles constitute a prime resource for the compilation of parallel corpora» [10]. Their contribute to the field of subtitle-corpus writing has resulted in the project *OPUS*, the largest subtitles parallel corpus, based on the collection of *Open Subtitle* database[1]. Project OPUS represents a fundamental resource for preliminary analysis on linguistic features, containing 1,689 texts from over 60 languages. The amount of the information provided is immense, with 2.6 billion sentences and over 17.2 billion tokens available for bilingual linguistic investigation.

It is important to point out that the validity of corpora based on subtitles is strongly linked to their written form, which represents a standardization of 'what is being said on screen', Tiedemann (2007) recalls the importance of studies in TV series and movies subtitles in the form of corpora, highlighting the strong connection between subtitles and their corresponding *source material* [11]. In addition, subtitles are valuable for corpus-analysis despite the apparent difference between subtitles language and natural conversation language, particularly regarding *pro-forms* and politeness markers. Levishina (2017) evidences that, while for natural conversation «reflects the Speaker's construal of the communicative situation and the relationship with the Hearer in terms

[1] http://opus.nlpl.eu/OpenSubtitles-v2016.php/.

of social distance, power and other parameters», in subtitles, on the other hand «the cognitive mechanism involved are much more complex» [12]. If it's true that one of the most striking differences between conversational language and subtitles language is in the lower presence of narrative elements in subtitles (Bednarek, 2011) [13]; on the contrary, lexicogrammar and pragmatic features have a similar incidence in both conversational and subtitles languages (Dose, 2014) [14]. Therefore, those features, which are fundamental in the computational linguistic survey, could represent the basis for linguistic analysis on intralingual subtitles corpora.

An important question that may arise is, to what extent (and in which areas consequently) could intralingual subtitles be more effective than interlingual, or, translating subtitles for linguistic consideration?

Caimi (2006) has underlined the importance of intralingual subtitles in the learning environment in connection with vocabulary formation and linguistic memory. One of the key features of intralingual subtitles (that distinguished intralingual and interlingual subtitles), is the significant correspondence between spoken text of the video source and the written text of subtitles. Caimi evidences that «if there is no biunique correspondence between spoken text and written text, comprehension is undermined, and students' feedback is exposed not only to phonological and orthographic inaccuracies but also to semantic confusion» [15]. Highlighting the key role of intralingual subtitles in vocabulary formation is the starting point for a more detailed linguistic analysis.

Another trait to be considered is that *collocations* are equally reported in intralingual subtitles as they appear in the spoken text of the source video. Collocation (grouping two or more words that usually occur together) is a fundamental branch of investigation for corpus linguistics and its relevance is the same for spoken and written texts.

Carstens (2016) emphasizes that understandability of complex linguistic features, «for example, the use of idioms collocations and subject specific vocabulary might be difficult for non-English speaking people to understand» and, therefore, «intralingual subtitles can be used […] for recognition and recall» [16]. Extract collocations data is possible through the building of POS-tagged corpora, allowing linguists and teachers to investigate the most common patterns in specific subtitle texts.

The observations given earlier might suggest that intralingual-subtitles corpora are only relevant for language classes. This is partly true, as the majority of enquiries results supportive for language learners. But there are more applications that, despite their lower exploration, can represent in the future a crucial progress in computational linguistics and AVT cross-studies.

In fact, one of the crucial aspects that must be considered in the building of subtitling corpora is the variety of *genres* of audiovisual products; building different corpora sorted by genre may highlight different communicative aspects of contemporary spoken language (like targeted audience, communicative situations, power and purpose), embracing sociolinguistic perspectives (Diaz-Cintas, Nikolić, 2017) [17] and discourse analysis as well (Cordella, 2006) [18]. For example, a teen-drama spectator expects a specific linguistic pattern, close to young people language and full of the so-called 'internet expressions'; it is important to *encode* key features in 'teen-English' which is not the same language spoken, for example, by an inspector during an interrogation with a murder suspect in a crime series. The research made for this paper will try to investigate what type of analysis can be made through intralingual-subtitles corpora.

3. BUILDING THE *CORPORA*

The research made for this paper aims to give an overall examination on how subtitles corpora are not only useful for translators in the form of parallel corpora, but also intralingual-subtitles corpora can be concrete tools for linguists and language teachers; particularly, their utility in the

general understanding of the linguistic structure of the original script, observing tendencies and collocations in contemporary English language.

The two series taken into consideration for this paper are not randomly selected from the wide range of Netflix 'Trending Now' list. They are different in genre and targeted audience, but the two have in common a sense of controversy of the contemporary values and are actual in their own ways. This sense of controversy and modernity is well established in the language used in the two shows: *Insatiable* is a teen-comic-drama that is not afraid to speak in a satirical and 'light' tone of something like *body and fat shaming*, describing in a fresh and surprisingly actual language the 'ideal redemption path' of a fat girl-becoming-skinny. The apparent discrimination and derision of a difficult and contemporary social problem got the attention (and often the critics) of viewers and journalists, who superficially paid no or little attention to the real social critique that the show wanted to provide [19]. The social critique is the *leitmotiv* of *Black Mirror*, a show that depicts the humanity as a voracious entity so hungry for innovation to have lost the faith in itself, relying only on the scientific progress and fiercely believing in the 'machines will overcome men' mantra. Themes like invasion of privacy, artificial wars, virtual-reality tortures and human slavery to technology are treated with so frightfully naturalness to scare and make the audience reflect to the problems of our society. Linguistic analysis on the two corpora will raise eventually thoughts and suggestions that can be matter of a sociolinguistics study on the two shows.

3.1 Materials

To analyze the written files of the subtitles in this paper, the corpora were made through *Sketch Engine*, a corpus manager and text analysis web-tool developed by Lexical Computing Limited[2]. Its 'user-friendly' interface and its various features made that choice proficient for the purpose of analysis. The subtitles are collected to have the same number of episodes considered for both shows (12), giving that the series have similar average length-per episode (45 minutes). This is the reason why, while for *Insatiable* the subtitles are taken for its first (and only at the time) season, for *Black Mirror* are taken into consideration the complete seasons three and four. The preference for the latest seasons of Black Mirror arises from the proximity of the years when the shows are produced: Black Mirror Season Three aired from the end of 2016 to the beginning of 2017, and its Season Four aired from the end of 2017 to the beginning of 2018 (Insatiable aired in the late summer of 2018). The corpora made with Sketch Engine are not the only taken in consideration for this paper: in order to have correct parameters, references and comparisons data are collected from spoken corpora only.

3.2 Lexical Analysis

For this section, the reference corpus is SBCSAE (Santa Barbara Corpus of Spoken American English), one of the largest corpora of contemporary spoken American English[3]. The first step is to examine the lexical density of the two corpora. Lexical density refers as the number of lexical (or content) words divided by the total number of words (Halliday, 1985) [20]. Lexical words are nouns, adjectives, verbs and adverbs. The general belief is that the lexical density of the spoken language is lower than the one in written texts (Stegen, 2007) [21]. The Insatiable Corpus has 33,582 lexical items out of 87,095. It is important for the purpose of analysis the normalization of the frequency to compare data with other bigger corpora. To do this, for this paper it will be considered the per million frequency. The Insatiable Corpus contains 385,579 words per million. This means that the lexical density in the Insatiable Corpus is 38,55%. As regards the Black Mirror Corpus, it has 24,569 lexical words out of 72,948. Per million words, it contains 336,801 lexical items, for a 33,68% of lexical density. SBCSAE has 331,303 lexical items per million (33,13%).

[2] http://sketchengine.academia.edu/.
[3] https://www.linguistics.ucsb.edu/research/santa-barbara-corpus/.

	Ins. Corpus	*B.M. Corpus*	*SBCSAE (Ref.)*
Nouns *(per mil.)*	112,050	100,619	197,099
Adj. *(per mil.)*	31,632	24,279	16,504
Verbs *(per mil.)*	178,230	152,355	82,001
Adv. *(per mil.)*	63,666	59,096	35,697
Lexical Density	38,55%	33,68%	33,13%

TABLE 1: Lexical items count per million and lexical density in the corpora.

This is perfectly in line with the studies of Ure (1971), who states that the lexical density of a spoken text is generally under 40% and drastically lower than the lexical density of a written text [22]. It is interesting noticing that, while for nouns the SBCSAE counts a significantly higher number of words than the other two corpora, in the other three categories this corpus has far less tokens and the percentage of lexical density is lower than the two corpora made for this paper. To be more precise, the following pie charts will explain the differences in the three corpora. Comparing the density of nouns, per million counts, the Insatiable Corpus represents 27,3% Black Mirror 24,5% and SBCSAE a large 48,2% of the total counts of nouns of the three corpora (409,768).

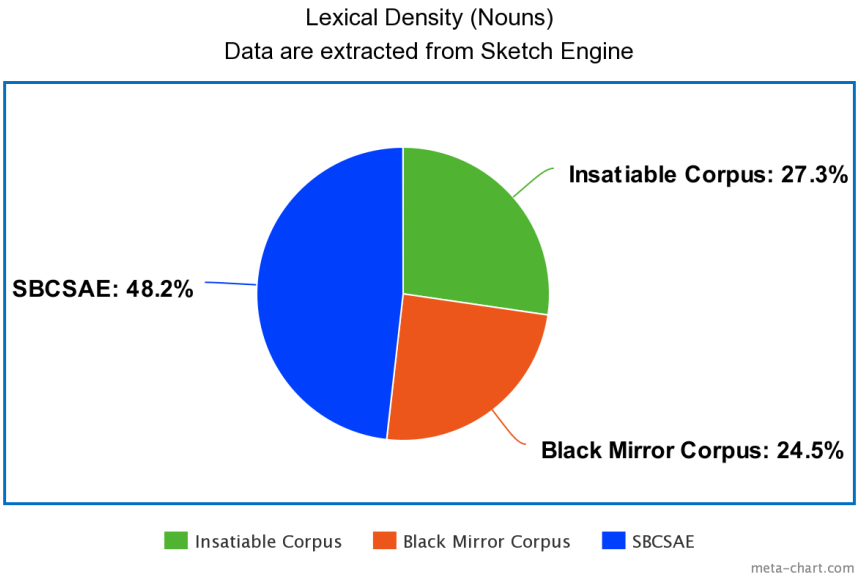

FIGURE 1: Comparing nouns density chart in percentage. Made with *Meta Chart*[4].

In the other three categories the data overturn. For adjectives, the Insatiable Corpus takes the 43,7%, the Black Mirror Corpus the 33,5 and the SBCSAE only the 22,8% (72,415 total tokens).

[4] https://www.meta-chart.com/.

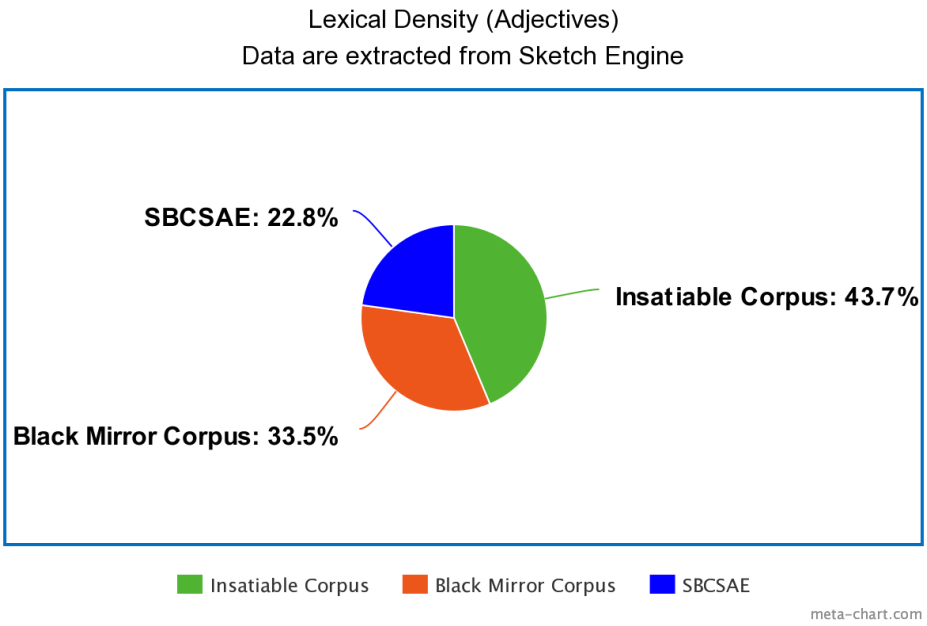

FIGURE 2: Comparing adjectives density chart in percentage. Made with *Meta Chart*.

Data are similar for verbs and adverbs. Insatiable Corpus verbs take 43,2%, Black Mirror Corpus 36,13% and SBCSAE 20,7% of 412,586 total verbs. As regards adverbs, Insatiable Corpus 40,2%, Black Mirror Corpus 37,3%, while SBCSAE has only the 22,5% of 158,459 adverbs.

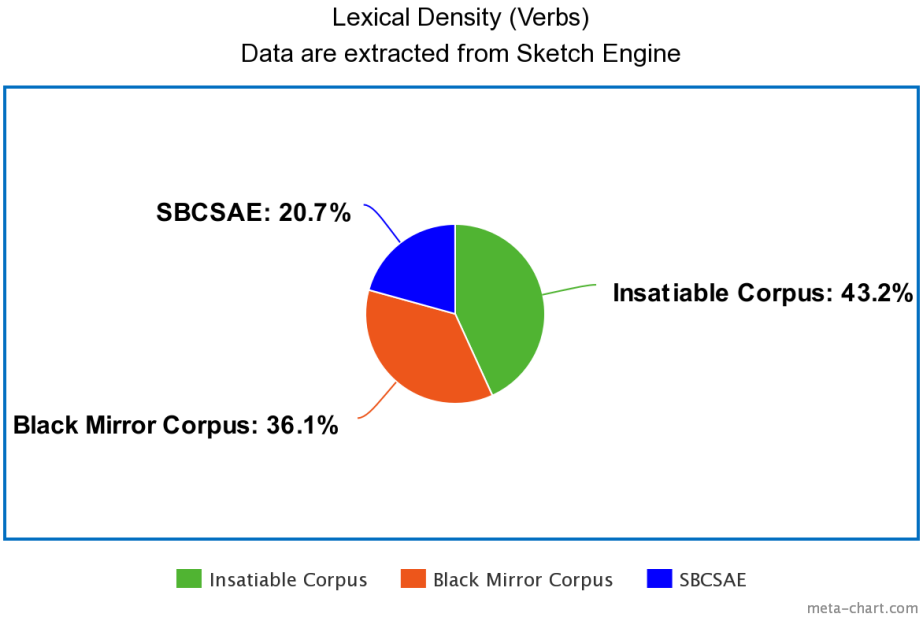

FIGURE 3: Comparing verbs density chart in percentage. Made with *Meta Chart*.

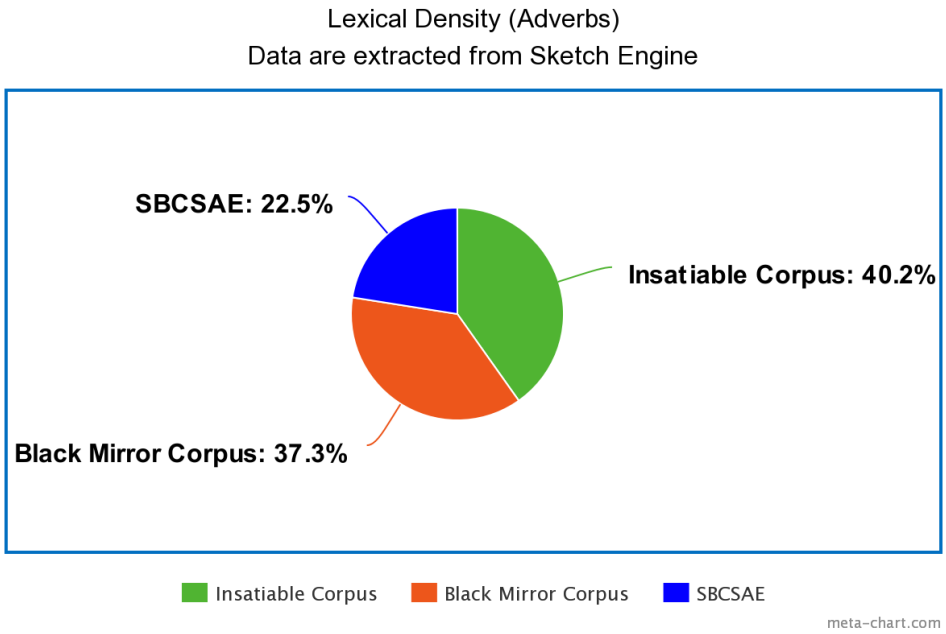

FIGURE 4: Comparing adverbs density chart in percentage. Made with *Meta Chart*.

Some final considerations for this section need to be made. It is clear that the lexical density of the two series corpora is similar, but a considerable difference is the variety of lexical elements occurring in them: *Insatiable*, as a teen-comic-drama, profoundly relies on exchanges and long discourses. Wordplays and puns are made with variations on adjectives: for example, one of the most visible linguistic features of Insatiable is the high frequency of adjectives starting with the letter 'B' used by one of the two protagonists to describe with not-uplifting epithets his rival, Bob Barnard. *Black Mirror* is an interesting series for any linguistics analysis, but it is fundamentally based on an imaginative style which has the primary objective in capturing the viewers with its incredible visual effects and only in a latter view the audience can catch the valuable linguistic patterns contained in each individual episode. Taking into account this fact, it is not surprising that Insatiable has a higher and more diverse lexical density than *Black Mirror*.

3.3 Word Frequency

Word frequency is how many times a word recurs in a single corpus. It is important again to normalize the results per million to have the right numbers for comparisons. The two tables listed below are the per-mil frequencies of words in the two corpora.

INSATIABLE CORPUS WORD FREQUENCY

Words	Frequency (Per Mil.)
I	41,690
You	34,743
To	19,581
The	17,325
A	16,281
It	13,479
And	12,492
N't	11,378
Me	10,207
That	10,138
Do	9,208
My	9,150

Was	8,588
What	7,761
Of	7,554
Is	7,038
In	6,429
For	6,085
We	5,568
Have	5,373

TABLE 2: Insatiable twenty most-frequent words chart (per mil.).
Data are extracted from *Sketch Engine*.

BLACK MIRROR CORPUS WORD FREQUENCY

Words	Frequency (Per Mil.)
I	37,765
You	37,745
It	27,043
The	26,525
A	19,351
To	17,836
That	15,365
N't	13,452
And	12,017
Do	11,040
Of	10,164
What	10,024
In	9,785
Is	9,526
No	9,287
On	9,187
Just	8,789
Me	7,812
This	7,692
We	7,673

TABLE 3: Black Mirror twenty most-frequent words chart (per mil.).
Data are extracted from *Sketch Engine*.

The pronouns (or adjectives) 'I' and 'You' are the two most frequent words in these two corpora. This is not surprising, giving that in the most of spoken corpora the two are at least in the most-frequent twenty words. In the SBCSAE, 'I' is the second most-frequent word and 'You' the fifth one (in the spoken section of *COCA*, the *Contemporary Corpus of American English*[5], 'I' and 'You' occupy in its frequency-chart, respectively, the eleventh and the fourteenth position).

It must be considered that in a TV series, dialogues imply a higher usage of these pronouns because of the dialogical nature of the scenes. Per million, in the Insatiable Corpus the word 'I' recurs 54,243 times, in the Black Mirror Corpus 37,765 times. As a reference, in the SBCSAE 'I' recurs 26,207 times and in the spoken section of the COCA 18,401 times per million. The word 'You' recurs 45,205 times per million in the Insatiable Corpus, 37,745 times per million in the Black Mirror Corpus, while in SBCSAE and COCA are relatively lower, respectively 19,723 and 18,001.

[5] https://www.english-corpora.org/coca/.

Two examples will be provided to show how it is possible to investigate the particular features of a TV Series with the 'Word Frequency' tool, one for both Insatiable and Black Mirror. The adjective 'fatty', a word used to create the nickname for the female protagonist of the show, 'Fatty Patty', recurs in the corpus 24 times, 359 times per million. In the spoken section of COCA, it recurs 170 times, only 1.4 per million words, and in the SBCSAE there are no records for this word. It is evident that this word is widely current in the Insatiable corpus, while in the other corpora it is an almost-non-existent word.

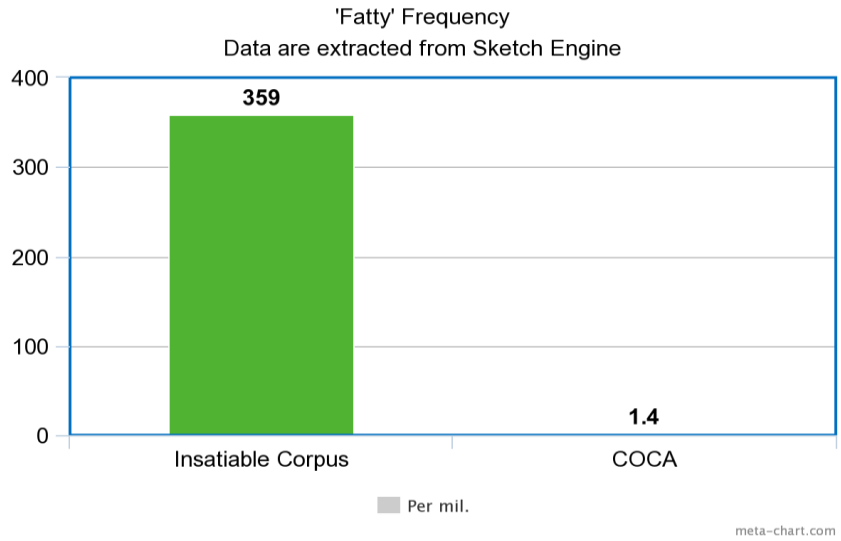

FIGURE 5: Fatty' frequency chart in the Insatiable Corpus and in the COCA. Made with *Meta Chart*.

To compare the word selected for the Black Mirror Corpus, the English Web Corpus seemed to be more effective. The word considered is 'device' in its contemporary use of 'electronic instrument'. In the Black Mirror Corpus this word appears 85 times, 1,694 per million. In the English Web Corpus, a higher number was to be expected and the word recurs 1,865,709 times. But the surprising data is that it recurs only 119 times per million, a recurrence drastically lower than the Series Corpus.

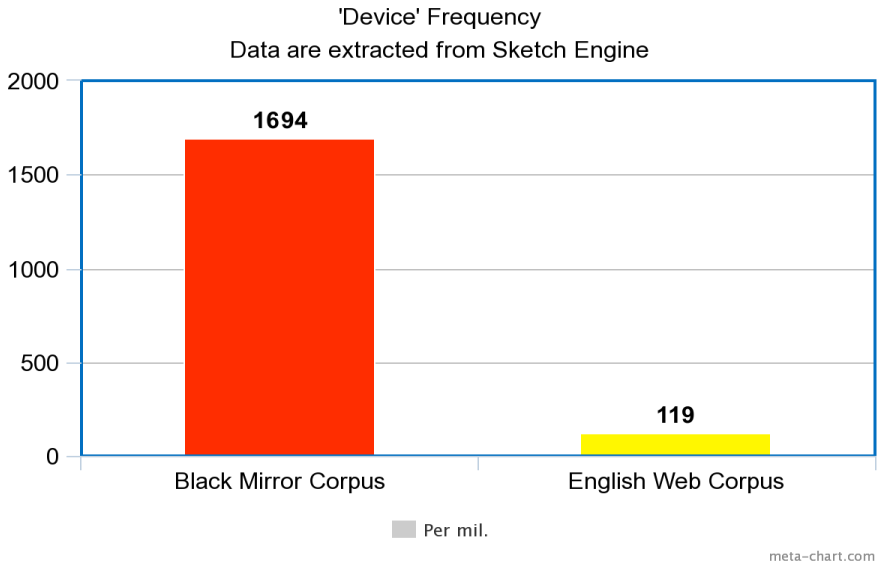

FIGURE 6: Device' frequency chart in the Black Mirror Corpus and in the English Web Corpus. Made with *Meta Chart*.

This kind of research is important to find recurrences and patterns more common in a specific series than in others.

3.4 Concordances

A concordance is a listing of each occurrence of a word (or pattern) in a text or corpus, presented with the words surrounding it. This is one of the most important tools for a translator or a linguist, because it can show how specific patterns behave in context. Because of its largeness, for the purposes of the paper it will be shown only a couple of concordances for each of the two corpora analyzed, the most salient ones. In the Insatiable series, the world of pageants and the maniacal research of the ideal beauty are two of the main themes, treated with a subtle irony and sarcasm by the authors of the show. The table below shows, in contrast, what are the words related with two *keywords in context* (*KWIC*), 'beautiful' and 'fat'. For the purpose of the analysis, the more significant ones are selected, according to general linguistic accepted rules, eliminating therefore conjunctions, prepositions, auxiliary verbs and other 'non-significant forms' (Luhn, 1959) [23].

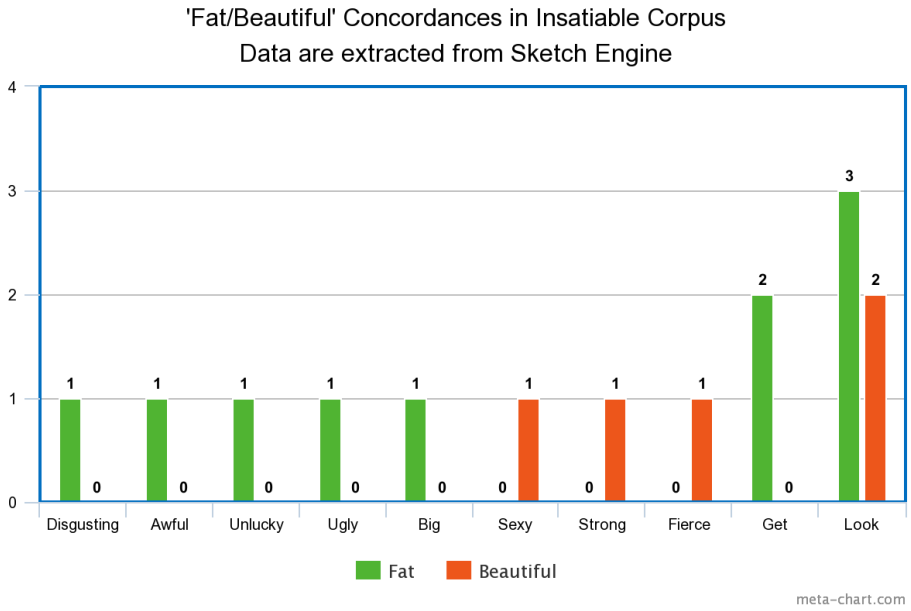

FIGURE 7: Concordances for words 'fat' and 'beautiful'. Made with *Meta Chart*.

The interesting fact is that the two adjectives have concordances that show how negative is the concept of a 'fat' person for the characters in the series. In fact, the word 'fat' comes with adjectives like 'disgusting', 'awful', 'unlucky', 'ugly'. The word 'beautiful' extremize even more the concept of 'obsession for beauty'. Adjectives like 'strong' and 'fierce' are not commonly used with the term 'beautiful'. The sense that apparently comes out of the usage of this adjective is that a 'beautiful' girl has a sort of a power that a 'fat' one doesn't have. This is one of the most striking socio-linguistic aspects which emerges from the analysis of the Insatiable Corpus.

As regards the Black Mirror Corpus, the concordances of the word 'God' are put in examination.

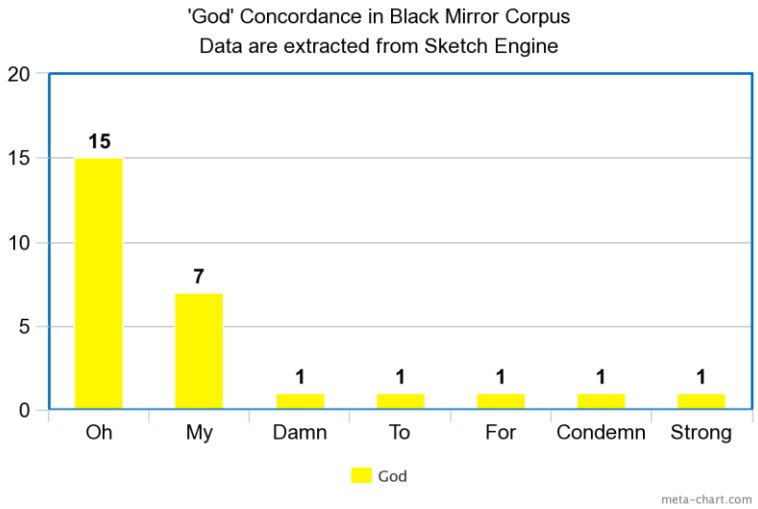

FIGURE 8: God' concordances in Black Mirror Corpus. Made with *Media Chart*.

The interesting fact is that, although no divinity or God-like entity is cited in the episodes of the series, the 'Oh, My God', or 'Oh, God' pattern is the most frequent one, showing the total controversy of the society described. The role of the translator in this case, too, is to find how the authors of the series wanted to communicate some linguistic messages and to adapt them in the best way they could.

4. DISCUSSION

This research has highlighted some of the linguistic features of two TV series with the help of corpus-based tools. While the two corpora have shown several traits in common with general spoken corpora, such as general lexical density in comparison with SBCSAE, various unique data appeared, starting from the specific lexical density of nouns and adjectives. The potential applications of the data shown above are countless: for example, comparing the relevance of lexical items in the audiovisual discourse with general spoken language, or finding what are the lexical categories which are affected by unusual frequency due to their specific genre.

Collocations and concordances revealed, even further, interesting results to a more detailed analysis. Extracting data and contextualizing them, could bring to the developing of collocation dictionaries based on intralingual subtitles. Integrating particular collocations and concordances with general spoken corpora, it is possible to highlight words, which are used in a singular way or within an uncommon pattern. For this reason, the words analyzed for the research were carefully selected to emphasize their particular use in the two subtitles transcriptions. However, analyzing more common patterns can, likewise, bring interesting results. It is important, finally, recall the importance of genre distinction; combine data from different types of audiovisual could lead to an erroneous interpretation of the results. For this reason, the two series were put in two different corpora. Socio-linguistic traits of the two series emerged within their context and become significant only when appropriately interpreted.

5. CONCLUSION

This paper has proposed a research work to evaluate a new perspective for computational linguistic studies, determining if it is possible to carry out a complete linguistic analysis on tendencies and features of audio-visual works. While previous research has stressed out the importance of corpora based on audiovisuals for specific purposes (Mustefa et al.) [24], and the crucial role of building of parallel corpora on intralingual subtitles for translators (Pavesi, 2018) [25], analysis on intralingual-subtitles corpora are still not substantial. The research made has shown what are the possible lines of investigation building corpora on intralingual subtitles. The examples displayed have only 'scratched the surface', as the opportunities for further

investigation are immense. Linguists can investigate tendencies on scripted language and distinguish between audio-visual language and natural spoken language in their particular features. Translators can rely on a more complete overview on the type of product they translate; knowing how English language operates in a specific audio-visual opera is crucial to translate and adapt it without 'falsify' the product translated (which is yet, unfortunately, a common feature of AVT). For teachers, building corpora based on audiovisuals can help students identify the most common patterns of spoken language, overcoming difficult sections in the understanding of a video in its source language, and learn strategies and approaches for the oral production.

It is possible, thereby, concluding that differentiate either intralingual and interlingual subtitles corpora, and genre-targeted subtitles corpora, could enrich the quality of studies in the field. In particular, the aim of the research carried on was collecting singularities of the two series and put them in context, making them relevant from a linguistic point of view. However, this is not the only exploration opportunity, as well as TV-series are not the only media that can be considered. The same analysis can be carried out to movies or videogames, or to institutional videos as well (politic debates, advertisings, documentaries, informative videos); each one of them can underline different type of communication, aspects and socio-cultural traits of the language.

The future challenge for researchers is to enhance the quality of intralingual-subtitles corpora, integrating them in the study of other spoken ones. Making connections, extracting data and developing new enquiries based on intralingual subtitles can improve the studies on computational approaches and can represent, imminently, one of the most valid tools for linguistic cross-studies.

6. REFERENCES

[1] S. Kübler. "Introduction to Corpus Linguistics." Internet: http://www.sfs.uni-tuebingen.de/~kuebler/rocoli/intro_corp_ling.pdf, Oct. 17, 2005 [Oct. 2, 2019].

[2] T. Schmidt, K. Wörner. (2009). "EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research." *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) Pragmatics* [On-line], 19(4), 565–582. Available: http://doi.org/10.1075/prag.19.4.06sch [Oct. 1, 2019].

[3] J. Newman. (2008). "Spoken corpora: Rationale and application." *Taiwan Journal of Linguistics* [On-line]. 6. 27-58. Available: http://doi.org/10.6519/TJL.2008.6%282%29.2 [Oct. 1, 2019].

[4] R. Baños, J. Díaz-Cintas. (2018). "Language and translation in film: dubbing and subtitling", in *The Routledge Handbook of Translation Studies and Linguistics*, 2nd ed., K. Malmkjær, Ed. London: Routledge, pp. 313-326.

[5] "The Official BBC Subtitle Guidelines." Internet: http://bbc.github.io/subtitle- guidelines, version 1. 1., May 7, 2018 [Sept. 6, 2019].

[6] K. Boonkit. (2010). "Enhancing the development of speaking skills for non-native speakers of English." *Procedia - Social and Behavioral Sciences* [On-line]. 2, 1305-1309. Available: http://doi.org/10.1016/j.sbspro.2010.03.191 [Oct. 1, 2019].

[7] W. Brown. "What Exactly Are Subtitles for the Deaf & Hard-of-Hearing (SDH)?". Internet: https://www.jbistudios.com/blog/what-exactly-are-subtitles-for-the-deaf-hard-of-hearing-sdh, Jul. 5, 2017 [Sept. 7, 2019].

[8] L. Bywood, M. Volk, M. Fishel, P. Georgakopoulou (2013). "Parallel subtitle corpora and their applications in machine translation and translatology." *Perspectives* [On-line]. 21(4), 595-610. Available: http://doi.org/10.1080/0907676X.2013.831920 [Oct. 2, 2019].

[9] S. A. Bird, J. N. Williams (2002). “The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling.” *Applied Psycholinguistics* [On-line]. 23. 509 - 533. Available: http://doi.org/10.1017/S0142716402004022 [Oct. 1, 2019].

[10] P. Lison, J. Tiedemann (2016). “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.” *LREC* [On-line]. Available: http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_paper.pdf [Sept. 30, 2019].

[11] J. Tiedemann (2007). “Building a multilingual parallel subtitle corpus.” *Proceedings of CLIN* [On-line]. 17. Available: http://pdfs.semanticscholar.org/0d0e/b34ab56f7b48b6d611b5d0767bc59ba8b9fc.pdf?_ga=2.131553159.969469141.1570094722-562003135.1570094722 [Oct. 1, 2019].

[12] N. Levshina (2017). “A Multivariate Study of T/V Forms in European Languages Based on a Parallel Corpus of Film Subtitles.” *Research in Languag e* [On-line], 15(2), 153-172. Available: http://doi.org/10.1515/rela-2017-0010 [Oct. 1, 2019].

[13] M. Bednarek (2011). “The Language of Fictional Television: A Case Study of the ‘Dramedy’ Gilmore Girls.” *English Text Construction* [On-line] 4(1). Available: http://dx.doi.org/10.1075/etc.4.1.04bed [Oct. 1, 2019].

[14] S. Dose (2014). “Describing and Teaching Spoken English: An Educational-Linguistic Study of Scripted Speech.” PhD Dissertation. *Giessen: Justus-Liebig-Universität Giessen* [On-line]. Available: http://d-nb.info/1068589043/34 [Oct. 1, 2019].

[15] A. Caimi (2006). “Audiovisual Translation and Language Learning: The Promotion of Intralingual Subtitles.” *The Journal of Specialized Translation* [On-line], 6. Available: https://www.jostrans.org/issue06/art_caimi.php [Oct. 2, 2019].

[16] A. Carstens (2016). “Translanguaging as a vehicle for L2 acquisition and L1 development: students’ perceptions.” *Language Matters* [On-line], 47(2), 203-222. Available: http://dx.doi.org/10.1080/10228195.2016.1153135 [Oct. 1, 2019].

[17] J. Díaz-Cintas, K. Nikolić. (2018). Fast-forwarding with audiovisual translation. (1st edition). *Bristol: Multilingual Matters* [On-line]. Available: http://doi.org/10.21832/DIAZ9368 [Oct. 2, 2019].

[18] M. Cordella (2017). “Discourse Analysis and the Subtitles of Documentaries: The Case of The Children Of Russia.” *ODISEA. Revista de estudios ingleses* [On-line]. Available: http://doi.org/10.25115/odisea.v0i7.143 [Oct. 2, 2019].

[19] F. Sturges. “Insatiable: there’s self-loathing for all in Netflix’s ‘fat-shaming’ teen comedy” From *The Guardian Online*. Internet: http://www.theguardian.com/tv-and-radio/2018/aug/10/insatiable-netflix-fat-shaming-debby-ryan. Aug. 10, 2018 [Aug. 31, 2019].

[20] M. A. K. Halliday. Spoken and written language. Geelong Vict.: *Deakin University*, 1985, 108 p.

[21] O. Stegen (2007). “Lexical Density in Oral versus Written Rangi Texts.” *SOAS Working Papers in Linguistics* [On-line], 15, 173-184. Available: https://www.researchgate.net/publication/315671445_lexical_density_in_oral_versus_written_rangi_texts [Oct. 2, 2019].

[22] J. Ure. “Lexical density and register differentiation.” In *Applications of Linguistics*, J.E. Perren, vol.1 no. 2. J.L.M. Trim Eds. London: Cambridge University Press, 1985, pp. 443-452.

[23] H. P. Luhn (1959). Keyword-in-context index for technical literature (KWIC index). (1st edition). [On-line]. Available: https://babel.hathitrust.org/cgi/pt?id=mdp.39015005511467 [Oct. 5, 2019].

[24] D. Mostefa, N. Moreau, K. Choukri, et al. (2007). “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms.” *Lang Resources & Evaluation* [On-line]. 41(389). Available: https://doi-org.pros.lib.unimi.it:2050/10.1007/s10579-007-9054-4 [Oct. 5, 2019].

[25] M. Pavesi. “Corpus-based audiovisual translation studies” in The Routledge Handbook of Audiovisual Translation, 1st edition. Luis Pérez-González Ed., *London: Routledge*, 2018. [On-line] pp.311-330. Available: https://www-taylorfrancis-com.pros.lib.unimi.it:2050/books/e/9781315717166/chapters/10.4324/9781315717166-20 [Oct. 5, 2019].