Serial Speakers: a Dataset of TV Series

Xavier Bost, Vincent Labatut, Georges Linares

Serial Speakers: a Dataset of TV Series

Xavier Bost, Vincent Labatut, Georges Linarès
Orkis, Laboratoire Informatique d'Avignon
13290 Aix-en-Provence, France, 84000 Avignon, France
{firstname.lastname}@univ-avignon.fr

Abstract

For over a decade, TV series have been drawing increasing interest, both from the audience and from various academic fields. But while most viewers are hooked on the continuous plots of TV serials, the few annotated datasets available to researchers focus on standalone episodes of classical TV series. We aim at filling this gap by providing the multimedia/speech processing communities with *Serial Speakers*, an annotated dataset of 161 episodes from three popular American TV serials: *Breaking Bad*, *Game of Thrones* and *House of Cards*. *Serial Speakers* is suitable both for investigating multimedia retrieval in realistic use case scenarios, and for addressing lower level speech related tasks in especially challenging conditions. We publicly release annotations for every speech turn (boundaries, speaker) and scene boundary, along with annotations for shot boundaries, recurring shots, and interacting speakers in a subset of episodes. Because of copyright restrictions, the textual content of the speech turns is encrypted in the public version of the dataset, but we provide the users with a simple online tool to recover the plain text from their own subtitle files.

Keywords: TV series, Multimedia retrieval, Speech processing.

Cite as:
Xavier Bost, Vincent Labatut, Georges Linarès.
Serial Speakers: a Dataset of TV Series.
12th International Conference on Language Resources and Evaluation (LREC 2020).

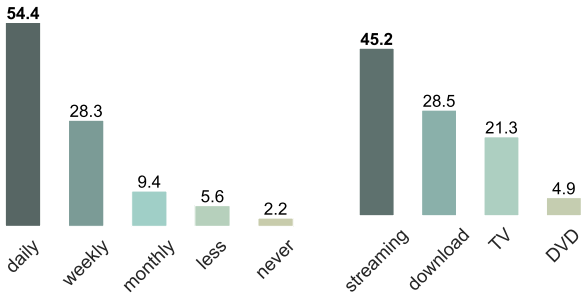
Note: This is a slightly extended version of the official LREC paper, including additional statistics for the final, eighth season of *Game of Thrones*, annotated after the paper was submitted.

1. Introduction

For over a decade now, TV series have been drawing increasing attention. In 2019, the final season of *Game of Thrones*, one of the most popular TV shows these past few years, has averaged 44.2 million viewers per episode; many TV series have huge communities of fans, resulting in numerous online crowdsourced resources, such as wikis[1], dedicated forums[2], and *YouTube* channels. Long dismissed as a minor genre by the critics, some recent TV series also received critical acclaim as a unique space of creativity, able to attract even renowned full-length movie directors, such as Jane Campion, David Fincher or Martin Scorsese. Nowadays, TV series have their own festivals[3]. For more than half of the people[4] we polled in the survey reproduced in (Bost, 2016), watching TV series is a daily occupation, as can be seen on Fig. 1a.

Such a success is probably related to the cultural changes caused by modern media: high-speed internet connections led to unprecedented viewing opportunities. As shown on Fig. 1b, television is no longer the main channel used to watch "TV" series: most of the time, streaming and downloading services are preferred to television.

Unlike television, streaming and downloading platforms give control to the user, not only over the contents he may want to watch, but also over the viewing frequency. As a consequence, the typical dozen of episodes a TV series season contains is often watched over a much shorter period of time than the usual two months it is being broadcast on television. As can be seen on Fig. 2a, for almost 80% of the people we polled, watching a TV series season (about 10 hours in average) never takes more than a few weeks. As a major consequence, TV series seasons, usually released once a year, are not watched in a continuous way.

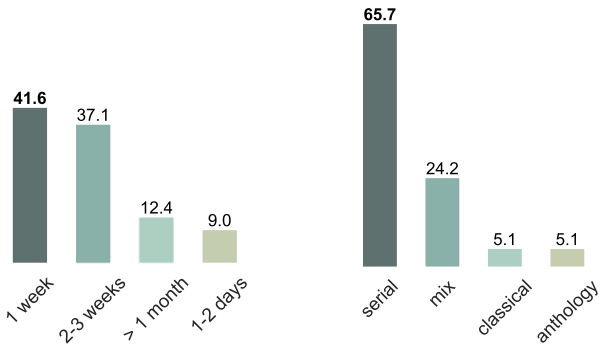


(a) Viewing frequency (%). (b) Viewing media (%).

Figure 1: TV series, viewing conditions.

For some types of TV series, discontinuous viewing is generally not a major issue. Classical TV series consist of self-contained episodes, only related with one another by a few recurring protagonists. Similarly, anthologies contain standalone units, either episodes (*e.g.* *The Twilight Zone*) or seasons (*e.g.* *True detective*), but without recurring characters. However, for TV *serials*, discontinuous viewing is

[1] gameofthrones.wikia.com/wiki/Game_of_Thrones_Wiki
[2] asoiaf.westeros.org/index.php?/forum
[3] In France, Series Mania.
[4] 194 individuals, mostly students from our university, aged 23.12 ± 5.73.



(a) Season viewing time. (b) Favorite genre.

Figure 2: TV series, season viewing time; favorite genre.

likely to be an issue: TV serials (*e.g. Game of Thrones*) are based on highly continuous plots, each episode and season being narratively related to the previous ones.

Yet, as reported on Fig. 2b, TV serials turn out to be much more popular than classical TV series: nearly 2/3 of the people we polled prefer TV serials to the other types, and 1/4 are more inclined to a mix between the classical and serial genres, each episode developing its own plot but also contributing to a secondary, continuous story.

As a consequence, viewers are likely to have forgotten to some extent the plot of TV serials when they are, at last, about to know what comes next: nearly 60% of the people we polled feel the need to remember the main events of the plot before viewing the new season of a TV serial. Such a situation, quite common, provides multimedia retrieval with remarkably realistic use cases.

A few works have been starting to explore multimedia retrieval for TV series. Tapaswi et al. (2014b) investigate ways of automatically building XKCD-style[5] visualizations of the plot of TV series episodes based on the interactions between onscreen characters. Ercolessi et al. (2012b) explore plot de-interlacing in TV series based on scene similarities. Bost et al. (2019) made use of automatic extractive summaries for re-engaging viewers with *Game of Thrones*' plot, a few weeks before the sixth season was released. Roy et al. (2014) and Tapaswi et al. (2014a) make use of crowdsourced plot synopses which, once aligned with video shots and/or transcripts, can support high-level, event-oriented search queries on TV series content.

Nonetheless, most of these works focus either on classical TV series, or on standalone episodes of TV serials. Due to the lack of annotated data, very few of them address the challenges related to the narrative continuity of TV serials. We aim at filling this gap by providing the multimedia/speech processing research communities with *Serial Speakers*, an annotated dataset focusing on three American TV serials: *Breaking Bad* (seasons 1–5 / 5), *Game of Thrones* (seasons 1–8 / 8), *House of Cards* (seasons 1–2 / 6). Besides multimedia retrieval, the annotations we provide make our dataset suitable for lower level tasks in challenging conditions (Subsection 3.1.). In this paper, we first describe the few existing related datasets, before detailing the main features of our own *Serial Speakers* dataset; we finally describe the tools we make available to the users for reproducing the copyrighted material of the dataset.

[5] xkcd.com/657

2. Related Works

These past ten years, a few commercial TV series have been annotated for various research purposes, and some of these annotations have been publicly released. We review here most of the TV shows that were annotated, along with the corresponding types of annotations, whenever publicly available.

Seinfeld (1989–1998) is an American TV *situational comedy (sitcom)*. Friedland et al. (2009) rely on acoustic events to design a navigation tool for browsing episodes publicly released during the ACM Multimedia 2009 Grand Challenge.

Buffy the Vampire Slayer (1997–2001) is an American *supernatural drama* TV series. This show was mostly used for character naming (Everingham et al., 2006), face tracking and identification (Bäuml et al., 2013), person identification (Bäuml et al., 2014), (Tapaswi et al., 2015b), story visualization (Tapaswi et al., 2014b), and plot synopses alignment (Tapaswi et al., 2014a)[6].

Ally McBeal (1997–2002) is an American *legal comedy-drama* TV series. The show was annotated for performing scene segmentation based on speaker diarization (Ercolessi et al., 2011) and speech recognition (Bredin, 2012), plot de-interlacing (Ercolessi et al., 2012b), and story visualization (Ercolessi et al., 2012a)[7].

Malcom in the Middle (2000–2006) is an American TV *sitcom*. Seven episodes were annotated for story de-interlacing (Ercolessi et al., 2012b) and visualization (Ercolessi et al., 2012a) purposes.

The Big Bang Theory (2007–2019) is also an American TV *sitcom*. Six episodes were annotated for the same visual tasks as those performed on *Buffy the Vampire Slayer*: face tracking and identification (Bäuml et al., 2013), person identification (Bäuml et al., 2014), (Tapaswi et al., 2015b), and story visualization (Tapaswi et al., 2014b). Tapaswi et al. (2012) also focus on speaker identification and provide audiovisual annotations for these six episodes[8]. In addition to these audiovisual annotations, Roy et al. (2014) publish in the TVD dataset other crowdsourced, linguistically oriented resources, such as manual transcripts, subtitles, episode outlines and textual summaries[9].

Game of Thrones (2011–2019) is an American *fantasy drama*. Tapaswi et al. (2014b) make use of annotated face tracks and face identities in the first season (10 episodes). In addition, Tapaswi et al. (2015a) provide the ground truth alignment between the first season of the TV series and the books it is based on[10]. For a subset of episodes, the TVD

[6] Visual (face tracks and identities) and linguistic (video alignment with plot synopses) annotations of the fifth season can be found at cvhci.anthropomatik.kit.edu/mtapaswi/projects-mma.html

[7] Annotations (scene/shot boundaries, speaker identity) of the first four episodes are available at herve.niderb.fr/data/ally_mcbeal

[8] cvhci.anthropomatik.kit.edu/mtapaswi/projectspersonid.html

[9] tvd.niderb.fr/

[10] cvhci.anthropomatik.kit.edu/mtapaswi/projectsbook_align.html

Show Season	Speech duration (ratio in %) BB	GOT	HOC	# speech turns BB	GOT	HOC	# speakers BB	GOT	HOC
1	02:01:19 (36)	03:32:55 (40)	04:50:12 (45)	4523	6973	11182	59	115	126
2	03:42:15 (38)	03:33:53 (41)	05:07:16 (48)	8853	7259	11633	86	127	167
3	03:42:04 (38)	03:30:01 (39)	_ (_)	7610	7117	_	85	115	_
4	03:38:08 (37)	03:11:28 (37)	_ (_)	7583	6694	_	70	119	_
5	04:40:03 (38)	02:55:32 (33)	_ (_)	10372	6226	_	92	121	_
6	_ (_)	02:48:48 (32)	_ (_)	_	5674	_	_	149	_
7	_ (_)	02:13:55 (32)	_ (_)	_	4526	_	_	66	_
8	_ (_)	01:27:17 (21)	_ (_)	_	3141	_	_	50	_
Total	17:43:53 (38)	23:13:52 (35)	09:57:29 (46)	38941	47610	22815	288	468	264

Table 1: Speech features.

dataset provides crowdsourced manual transcripts, subtitles, episode outlines and textual summaries.
As can be seen, many of these annotations target vision-related tasks. Furthermore, little attention has been paid to TV serials and their continuous plots, usually spanning several seasons. Instead, standalone episodes of sitcoms are overrepresented. And finally, even when annotators focus on TV serials (*Game of Thrones*), the annotations are never provided for more than a single season. Similar to the computer vision ACCIO dataset for the series of *Harry Potter* movies (Ghaleb et al., 2015), our *Serial Speakers* dataset aims in contrast at providing annotations of several seasons of TV serials, in order to address both the realistic multimedia retrieval use cases we detailed in Section 1., and lower level speech processing tasks in unusual, challenging conditions.

Show Season	Duration (# episodes) BB	GOT	HOC
1	05:32:44 (7)	08:58:28 (10)	10:48:15 (13)
2	09:51:08 (13)	08:41:56 (10)	10:37:10 (13)
3	09:49:40 (13)	08:52:04 (10)	_ (_)
4	09:46:16 (13)	08:41:05 (10)	_ (_)
5	12:15:36 (16)	08:56:50 (10)	_ (_)
6	_ (_)	08:55:43 (10)	_ (_)
7	_ (_)	06:58:54 (7)	_ (_)
8	_ (_)	06:48:31 (6)	_ (_)
Total	47:15:26 (62)	66:53:34 (73)	21:25:26 (26)

Table 2: Duration of the video recordings.

3. Description of the Dataset

Our *Serial Speakers* dataset consists of 161 episodes from three popular TV serials:

Breaking Bad (denoted hereafter BB), released between 2008 and 2013, is categorized on *Wikipedia* as a *crime drama*, *contemporary western* and a *black comedy*. We annotated 62 episodes (seasons 1–5) out of 62.

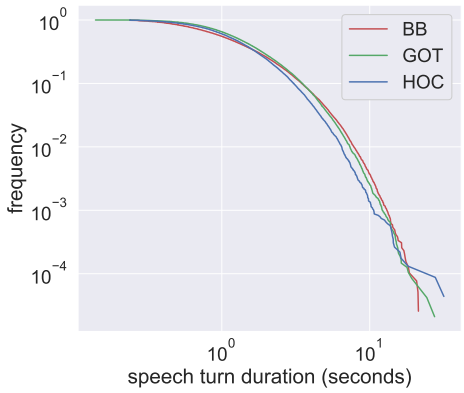
Game of Thrones (GOT) has been introduced above in Section 2. We annotated 73 episodes (seasons 1–8) out of 73.

House of Cards (HOC) is a *political drama*, released between 2013 and 2018. We annotated 26 episodes (seasons 1–2) out of 73.
Overall, the total duration of the video recordings amounts to $\simeq$ 135 hours (135:34:27). Table 2 details for every season of each of the three TV serials the duration of the video recordings, expressed in "HH:MM:SS", along with the corresponding number of episodes (in parentheses).

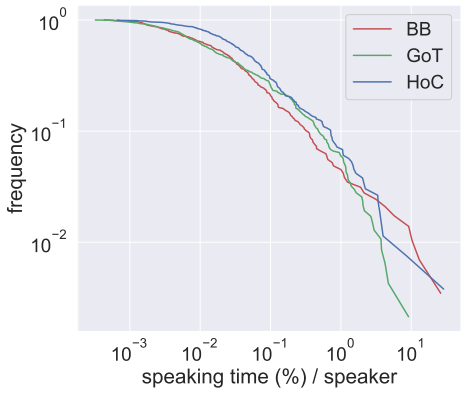
3.1. Speech Turns

As in any full-length movie, speech is ubiquitous in TV serials. As reported in Table 1, speech coverage in our dataset ranges from 35% to 46% of the video duration, depending on the TV series, for a total amount of about 51 hours. As can be seen, speech coverage is much more important (46%) in HOC than in BB and GOT (respectively 38% and 35%). As a political drama, HOC is definitely speech oriented, while the other two series also contain action scenes. Interestingly, speech coverage in GOT tends to decrease over the 8 seasons, especially from the fifth one. The first seasons turn out to be relatively faithful to the book series they are based on, while the last ones tend to depart from the original novel. Moreover, with increasing financial means, GOT progressively moved to a pure fantasy drama, with more action scenes.
The basic speech units we consider in our dataset are *speech turns*, graphically signaled as sentences by ending punctuation signs. Unlike speaker turns, two consecutive speech turns may originate in the same speaker.

Boundaries. The boundaries (starting and ending points) of every speech turn are annotated. During the annotation process, speech turns were first based on raw subtitles, as retrieved by applying a standard OCR tool to the commercial DVDs. Nonetheless, subtitles do not always correspond to speech turns in a one-to-one way: long speech turns usually span several consecutive subtitles; conversely, a single subtitle may contain several speech turns, especially in case of fast speaker change. We then applied simple merging/splitting rules to recover the full speech turns from the subtitles, before refining their boundaries by using the forced alignment tool described in (McAuliffe et al., 2017). The resulting boundaries were systematically inspected and



(a) Speech turns, duration distribution.



(b) Speaking time distribution.

Figure 3: Speech turns duration and speaking time/speaker.

manually adjusted whenever necessary. Such annotations make our dataset suitable for the *speech/voice activity detection* task.

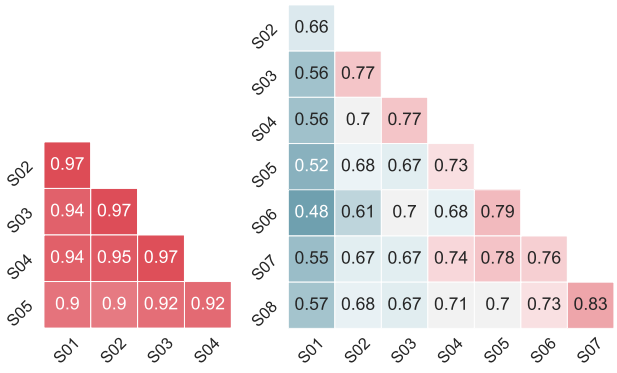
Overall, as reported in Table 1, the dataset contains 109,366 speech turns. Speech turns are relatively short: the median speech turn duration amounts to 1.3 seconds for GOT, 1.2 for HOC, and only 1.1 for BB.

As can be seen on Fig. 3a, the statistical distribution of the speech turns duration, here plotted on a log-log scale as a complementary cumulative distribution function, seems to exhibit a heavy tail in all three cases. This is confirmed more objectively by applying the statistical testing procedure proposed by Clauset et al. (2009), which shows these distributions follow power laws. This indicates that the distribution is dominated by very short segments, but that there is a non-negligible proportion of very long segments, too. It also reveals that the mean is not an appropriate statistic to describe this distribution.

Speakers. By definition, every speech turn is uttered by a single speaker. We manually annotated every speech turn with the name of the corresponding speaking character, as credited in the cast list of each TV series episode. A small fraction of the speech segments (BB: 1.6%, GOT: 3%, HOC: 2.2%) were left as unidentified ("unknown" speaker). In the rare cases of two partially overlapping speech turns, we decided to cut off the first one at the exact starting point of the second one to preserve as much as possible its purity.

Overall, as can be seen in Table 1, 288 speakers were identified in BB, 468 in GOT and 264 in HOC. With an average speaking time of 132 seconds by speaker, HOC contains more speakers than GOT (175 seconds/speaker), which in turn contains more speakers than BB (218 seconds/speaker).

Fig. 3b shows the distribution of the speaking time (expressed in percentage of the total speech time) for all speakers, again plotted on a log-log scale as a complementary cumulative distribution function. Once again, the speaking time of each speaker seems to follow a heavy-tailed distribution, with a few ubiquitous speakers and lots of barely speaking characters. This is confirmed through the same procedure as before, which identifies three power laws. If we consider that speaking time captures the strength of social interactions (soliloquies aside), this is consistent with results previously published for other types of weighted social networks (Li and Chen, 2003; Barthélemy et al., 2005). Nonetheless, as can be seen on the figure, the main speakers of GOT are not as ubiquitous as the major ones in the other two series: while the five main protagonists of BB and HOC respectively accumulate 64.3 and 48.6% of the total speech time, the five main characters of GOT "only" accumulate 25.6%. Indeed, GOT's plot, based on a choral novel, is split into multiple storylines, each centered on one major protagonist.

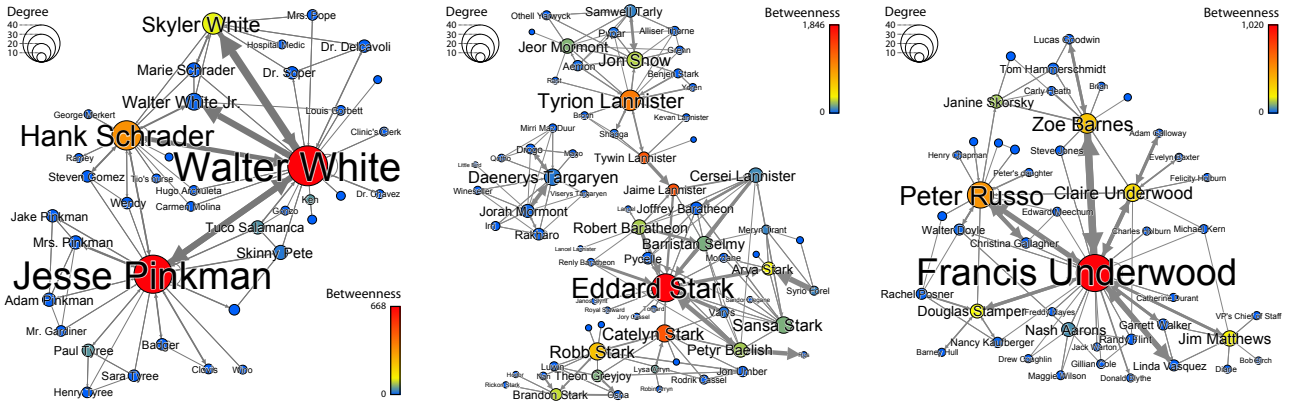


(a) BB (b) GOT

Figure 4: Speakers correlation across seasons.

Moreover, even major, recurring characters of TV serials are not always uniformly represented over time. Fig. 4 depicts the lower part of correlation matrices computed between the speakers involved in every season of BB (Fig. 4a) and GOT (Fig. 4b): the distribution of the relative speaking time of every speaker in each season is first computed, before the Pearson correlation coefficient is calculated between every pair of season distribution.

As can be seen, the situation is very contrasted, depending on the TV serial. Whereas the major speakers of BB remain quite the same over all five seasons (correlation coefficients close to 1, except for the very last, fifth one, with a few entering new characters), GOT exhibits quite lower correlation coefficients. For instance, the main speakers involved in the first season turn out to be quite different from the speakers involved in the other ones (average correlation coefficient



(a) BB. (b) GOT. (c) HOC.

Figure 5: Conversational networks extracted from the annotated episodes. Vertex size and color represent degree and betweenness, respectively.

with the other seasons only amounting to 0.56 ± 0.05). Indeed, GOT is known for numerous, shocking deaths of major characters[11]. Moreover, GOT's narrative usually focuses alternatively on each of its multiple storylines, but may postpone some of them for an unpredictable time, resulting in uneven speaker involvement over seasons. Fig. 6 depicts the relative speaking time in every season of the 12 most active speakers of GOT. As can be seen, some characters are barely present in some seasons, for instance, Jon (rank #4) in Season 2, or even absent, like Tywin (rank #12) in Seasons 5–8.

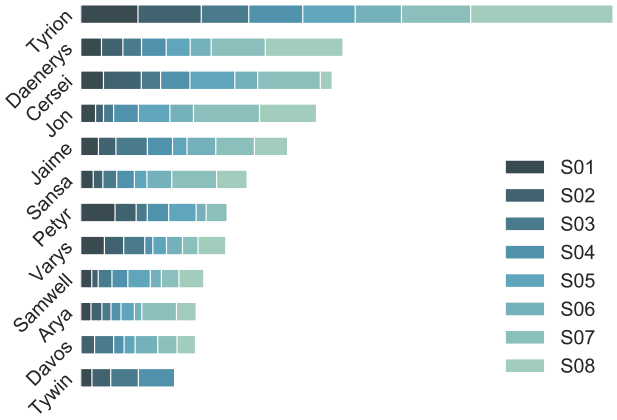


Figure 6: Relative speaking time over every season of the top-12 speakers of GOT.

Furthermore, as can be noticed on Fig. 6, the relative involvement of most of these 12 protagonists in Seasons 7–8 is much more important than in the other ones: indeed, Seasons 7–8 are centered on fewer speakers (respectively 66 and 50 *vs.* 124.3 ± 11.8 in average in the first six ones). Speaker annotations make our dataset suitable for the speaker diarization/recognition tasks, but in especially challenging conditions: first, and as stated in (Bredin and Gelly, 2016), the usual 2-second assumption made for the speech turns by most of the state-of-the-art speaker diarization systems does no longer stand. Second, the high number of speakers involved in TV serials, along with the way their utterances are distributed over time, make one-step approaches particularly difficult. In such conditions, multi-stage approaches should be more effective (Tran et al., 2011). Besides, as noted in (Bredin and Gelly, 2016), the spontaneous nature of the interactions, the usual background music and sound effects heavily hurt the performance of standard speaker diarization/recognition systems (Clément et al., 2011).

Textual content. Though not provided with the annotated dataset for obvious copyright reasons[12], the textual content of every speech turn has been revised, based on the output of the OCR tool we used to retrieve the subtitles. In particular, we restored a few missing words, mostly for BB, the subtitles sometimes containing some deletions.

BB contains 229,004 tokens (word occurrences) and 10,152 types (unique words); GOT 317,840 tokens and 9,275 types; and HOC 153,846 tokens and 8,508 types.

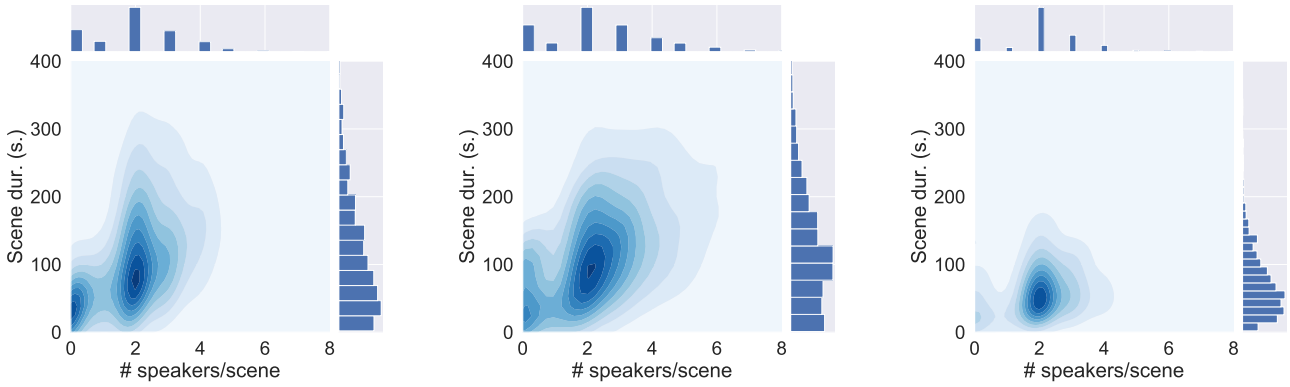
As the number of tokens vary dramatically from one TV serial to the other, we used the length-independent MTLD measure (McCarthy and Jarvis, 2010) to assess the lexical diversity of the three TV serials. With a value of 88.2 (threshold set to 0.72), the vocabulary in HOC turns out to be richer than in GOT (69.6) and BB (64.5). More speech oriented, HOC also turns out to exhibit more lexical diversity than the other two series.

3.2. Interacting Speakers

In a subset of episodes, the addressees of every speech turn have been annotated. Trivial within two-speaker sequences, such a task, even for annotators, turns out to be especially challenging in more complex conditions: most of the time, the addressees have to be inferred both from visual clues and from the semantic content of the interaction. In soliloquies (not rare in HOC), the addressee field was left empty.

[11] See got.show, for an attempt to automatically predict the characters who are the most likely to die next.

[12] Instead, we provide the users with online tools for recovering the textual content of the dataset from external subtitle files. See Section 4. for a description.



(a) BB. (b) GOT. (c) HOC.

Figure 7: # speakers/scene *vs.* scene duration.

Not frequently addressed alone, the task of determining the interacting speakers is nonetheless a prerequisite for social network-based approaches of fiction work analysis, which generally lack annotated data to intrinsically assess the interactions they assume (Labatut and Bost, 2019). Moreover, speaker diarization/recognition on the one hand, detection of interaction patterns on the other hand, could probably benefit from one another and be performed jointly. As an example, Fig. 5 shows the conversational networks based on the annotated episodes for each serial. The vertex sizes match their degree, while their color corresponds to their betweenness centrality. This clearly highlights the obvious main characters such as Walter White (BB) or Francis Underwood (HOC); but also more secondary characters that have very specific roles narrative-wise, *e.g.* Jaime Lannister who acts as a bridge between two groups of characters corresponding to two distinct narrative arcs. This illustrates the interest of leveraging the social network of characters when dealing with narrative-related tasks.

3.3. Shot Boundaries

Besides speech oriented annotations, the *Serial Speakers* dataset contains a few visual annotations. For the first season of each of the three TV series, we manually annotated shot boundaries. A video shot, as stated in (Koprinska and Carrato, 2001), is defined as an "unbroken sequence of frames taken from one camera". Transitions between video shots can be gradual (fade-in/fade-out), or abrupt ones (cuts). Most of the shot transitions in our dataset are simple cuts.

The first seasons of BB, GOT, and HOC respectively contain 4,416, 9,375 and 8,783 shots, with and average duration of 4.5, 3.4 and 4.4 seconds. Action scenes in GOT are likely to be responsible for shorter shots in average.

Shot boundary detection is nowadays well performed, especially when consecutive shots are abruptly transitioning from one another. As a consequence, it is rarely addressed for itself, but as a preliminary task for more complex ones.

3.4. Recurring Shots

Shots rarely occur only once in edited video streams: in average, a shot occurs 10 times in BB, 15.2 in GOT and 17.7 in HOC. Most of the time, dialogue scenes are responsible for such shot recurrence. As can be seen on Fig. 8, within dialogue scenes, the camera typically alternates between the interacting characters, resulting in recurring, possibly alternating, shots.

Figure 8: Example of two alternating recurring shots.

We manually annotated such recurring shots, based on similar framing, in the first season of the three TV series. As stated in (Yeung et al., 1998), recurring shots usually capture interactions between characters. Relatively easy to cluster automatically, recurring shots are especially useful to multimodal approaches of speaker diarization (Bost et al., 2015). Besides, recurring shots often result in complex interaction patterns, denoted *logical story units* in (Hanjalic et al., 1999). Such patterns are suitable for supporting local speaker diarization approaches (Bost and Linares, 2014), or for providing extractive summaries with consistent subsequences (Bost et al., 2019).

3.5. Scene Boundaries

Scenes are the longest units we annotated in our dataset. As required by the rule of the three unities classically prescribed for dramas, a scene in a movie is defined as a homogeneous sequence of actions occurring at the same place, within a continuous period of time.

Though providing annotators with general guidelines, such a definition leaves space for interpretation, and some subjective choices still have to be made to annotate scene boundaries.

First, temporal discontinuity is not always obvious to address: temporal ellipses often correspond to new scenes, but sometimes, especially when short, they hardly break the narrative continuity of the scene.

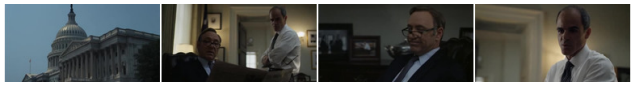


Figure 9: Long shot opening a scene.

Show / Season	Speech turns–Scenes BB	GOT	HOC	Shots BB	GOT	HOC	Interlocutors BB	GOT	HOC
1	✓	✓	✓	✓	✓	✓	4, 6	3, 7, 8	1, 7, 11
2	✓	✓	✓	✗	✗	✗	3, 4	✗	✗
3	✓	✓	–	✗	✗	–	✗	✗	–
4	✓	✓	–	✗	✗	–	✗	✗	–
5	✓	✓	–	✗	✗	–	✗	✗	–
6	–	✓	–	–	✗	–	–	✗	–
7	–	✓	–	–	✗	–	–	✗	–
8	–	✓	–	–	✗	–	–	✗	–

Table 3: Annotation overview.

Second, as shown on Fig. 9, scenes often open with long shots that show the place of the upcoming scene. Though there is no, strictly speaking, spatial continuity between the first shot and the following ones, they obviously belong to the same scene, and should be annotated as such.

Finally, action homogeneity may also be tricky to assess. For instance, a phone call within a scene may interrupt an ongoing dialogue, resulting in a new phone conversation with another character, and possibly in a new action unit. In such cases, we generally inserted a new scene to capture the interrupting event, but other conventions could have been followed. Indeed, the choice of scene granularity remains highly dependent on the use case the annotators have in mind for annotating such data: special attention to speaker interactions would for instance invite to introduce more frequent scene boundaries.

Overall, BB contains 1,337 scenes, with an average duration of 127.1 seconds; GOT 1,813 scenes (avg. duration of 132.6 seconds); HOC 1,048 scenes (avg. duration of 73.2 seconds). Once again, HOC contrasts with the two other series, with many short scenes.

Fig. 7 shows the joint distribution of the number of speakers by scene and the duration of the scene. For visualization purposes, the joint distribution is plotted as a continuous bivariate function, as fitted by applying kernel density estimate.

As can be seen from the marginal distribution represented horizontally above each plot, the number of speakers in each scene remains quite low: 2 in average in BB, 2.1 in HOC, and a bit more (2.4) in GOT. Besides, the number of characters in each scene, except maybe in GOT, is not clearly correlated with its duration. Moreover, some short scenes surprisingly do not contain any speaking character: most of them correspond to the opening and closing sequences of each episode. Finally, the short scenes of HOC generally contain two speakers.

Table 3 provides an overview of the annotated parts of the *Serial Speakers* dataset, along with the corresponding types of annotations. In the table, "Speech turns" stand for the annotation of the speech turns (boundaries, speaker, text); "Scenes" for the annotation of the scene boundaries; "Shots" for the annotation of the recurring shots and shot boundaries; and "Interlocutors" for the annotation of the interacting speakers[13].

4. Text Recovering Procedure

Due to copyright restrictions, the published annotation files do not reproduce the textual content of the speech turns. Instead, the textual content is encrypted in the public version of the *Serial Speakers* dataset, and we provide the users with a simple toolkit to recover the original text from their own subtitle files[14].

Indeed, the overlap between the textual content of our dataset and the subtitle files is likely to be large: compared to the annotated text, subtitles may contain either insertions (formatting tags, sound effect captions, mentions of speaking characters when not present onscreen), or some deletions (sentence compression), but very few substitutions. Every word in the transcript, if not deleted, generally has the exact same form in the subtitles. As a consequence, the original word sequence can be recovered from the subtitles. Our text recovering algorithm first encrypts the tokens found in the subtitle files provided by the user, before matching the resulting sequence with the original encrypted token sequence. The general procedure we detail below is likely to be of some help to annotators of other movie datasets with similar copyrighted material.

4.1. Text Encryption

For the encryption step, we used truncated hash functions because of the following desirable properties: deterministic, hash functions ensure that identical words are encrypted in the same way in the original text and in the subtitles; they do not reveal information about the original content, allowing the public version of our dataset to comply with the copyright restrictions; they are efficient enough to quickly process the thousands of word types contained in the subtitles; moreover, once truncated, hash functions result in collisions, able to prevent simple dictionary attacks. Indeed, the main requirement in our case is only to prevent collisions from occurring too close from each other: even if two different words were encrypted in the same way, they

[13] The annotation files are available online at: doi.org/10.6084/m9.figshare.3471839

[14] The toolkit is available online at: github.com/bostxavier/Serial-Speakers

would unlikely be close enough to result in ambiguous subsequences.
In the public version of our dataset, we compute the first three digits of the SHA-256 hash function of all of the tokens (including punctuation signs) and the exact same encryption scheme is applied to the subtitle files, as provided by the users, resulting in two encrypted token sequences for every episode of the three TV series.

4.2. Subtitle Alignment

We then apply to the two encrypted token sequences the *Python* `Difflib` sequence matching algorithm[15], built upon the approach detailed in (Ratcliff and Metzener, 1988).
Once aligned with the encrypted subtitle sequence, the tokens of the dataset are decryted by retrieving from the subtitles the original words.

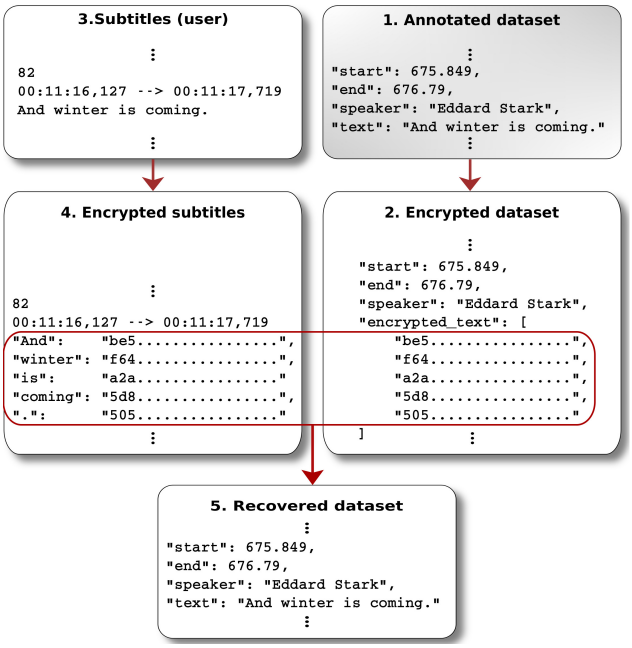


Figure 10: Text recovering procedure.

The whole text recovering procedure is summarized on Fig. 10. The annotated dataset with clear text, materialized by the gray box (Box 1) on the figure, is not publicly available. Instead, in the public annotations, the text is encrypted (Box 2). In order to recover the text, the user has to provide h(is/er) own subtitle files (Box 3), which are encrypted by our tool in the same way as the original dataset text (Box 4); the resulting encrypted token sequence is matched with the corresponding token sequence of speech turns (red frame on the figure), before the text of the speech turns is recovered from the subtitle words (Box 5).

4.3. Experiments and Results

In order to assess the text recovering procedure, we automatically recovered the textual content from external, publicly available subtitle files, and compared it to the annotated text. Table 4 reports in percentage for each of the three series the average error rates by episode, both computed at the word level (word error rate, denoted WER in the table) and at the sentence level (sentence error rate, denoted SER). In addition, we reported for every episode the average number of reference tokens (denoted *# tokens*), and the average number of insertions, deletions, and substitutions in the reference word sequence (respectively denoted *Ins*, *Del*, *Sub*). Because of possibly inconsistent punctuation conventions between the annotated and subtitle text, we systematically removed the punctuation signs from both sequences before computing the error rates.

Show	WER	SER	# tokens	Ins	Del	Sub
BB	1.6	4.6	3699.8	0.2	53.1	4.1
GOT	0.4	1.2	4353.2	0.1	13.8	1.9
HOC	0.2	0.7	5918.0	0.1	7.4	2.3

Table 4: Text recovering: avg. error rates (%) / episode.

As can be seen, the average error rates remain remarkably low: the word error rate amounts to less than 1% in average. The sentence error rate also remains quite low: about 1% for GOT and HOC, and a bit higher (4.6%) for BB. As can be seen in the right part of the table, deletions are responsible for most of the errors, especially in BB: as noted in Subsection 3.1., we restored the words missing in the subtitles when annotating the textual content of the speech turns. Such missing words turn out to be relatively frequent in BB, which can in part explain the higher number of deletions ($\simeq$ 53 deleted words in average out of $\simeq$ 3,700). Moreover, truncating the hash function to the first three digits does not hurt the performance of the text recovering procedure, while preventing simple dictionary attacks: the exact same error rates (not reported in the table) are obtained when keeping the full hash (64 hexadecimal digits).
In order to allow the user to quickly inspect and edit the differences between the annotated text and the subtitles, our tool inserts in the recovered dataset an empty tag <> at the location of deleted reference tokens. Similarly, we signal every substituted token with an enclosing tag (*e.g.* <Why>). As will be seen when using the toolkit, most of the differences come from different punctuation/quotation conventions between the annotation and subtitle files, and rarely impact the vocabulary or the semantics.
The whole recovering process turns out to be fast: 8.3 seconds for GOT (73 episodes) on a personal laptop (Intel Xeon-E3-v5 CPU); 6.73 for BB (62 episodes); 4.41 for HOC (26 episodes). We tried to keep the toolkit as simple as possible, with a single text recovering *Python* script with few dependencies.

5. Conclusion and Perspectives

In this work, we described *Serial Speakers*, a dataset of 161 annotated episodes from three popular TV serials, *Breaking Bad* (62 annotated episodes), *Game of Thrones* (73), and *House of Cards* (26). *Serial Speakers* is suitable for addressing both high level multimedia retrieval tasks in real world scenarios, and lower level speech processing tasks in challenging conditions. The boundaries, speaker and

[15] docs.python.org/3/library/difflib.html

textual content of every speech turn, along with all scene boundaries, have been manually annotated for the whole set of episodes; the shot boundaries and recurring shots for the first season of each of the three series; and the interacting speakers for a subset of 10 episodes. We also detailed the simple text recovering tool we made available to the users, potentially helpful to annotators of other datasets facing similar copyright issues.

As future work, we will first consider including the face tracks/identities provided for the first season of GOT in (Tapaswi et al., 2015a), but these face tracks, automatically generated, would need manual checking before publication. Furthermore, we plan to investigate more flexible text encryption schemes: due to the uniqueness property, hash functions, even truncated, are not tolerant to spelling/OCR errors in the subtitles. Though the correct word is generally recovered from the surrounding tokens, it would be worth investigating encryption functions that would preserve the similarity between simple variations of the same token.

6. Acknowledgements

This work was partially supported by the Research Federation *Agorantic* FR 3621, Avignon University.

7. Bibliographical References

Barthélemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. (2005). Characterization and modeling of weighted networks. *Physica A*, 346(1-2):34–43.

Bäuml, M., Tapaswi, M., and Stiefelhagen, R. (2013). Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609.

Bäuml, M., Tapaswi, M., and Stiefelhagen, R. (2014). A time pooled track kernel for person identification. In *11th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 7–12.

Bost, X. and Linares, G. (2014). Constrained speaker diarization of tv series based on visual patterns. In *IEEE Spoken Language Technology Workshop*, pages 390–395.

Bost, X., Linarès, G., and Gueye, S. (2015). Audiovisual speaker diarization of tv series. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4799–4803. IEEE.

Bost, X., Gueye, S., Labatut, V., Larson, M., Linarès, G., Malinas, D., and Roth, R. (2019). Remembering winter was coming. *Multimedia Tools and Applications*, 78(24):35373–35399, Dec.

Bost, X. (2016). *A storytelling machine? Automatic video summarization: the case of TV series*. Ph.D. thesis.

Bredin, H. and Gelly, G. (2016). Improving speaker diarization of tv series using talking-face detection and clustering. In *24th ACM international conference on Multimedia*, pages 157–161.

Bredin, H. (2012). Segmentation of tv shows into scenes using speaker diarization and speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2377–2380.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Clément, P., Bazillon, T., and Fredouille, C. (2011). Speaker diarization of heterogeneous web video files: A preliminary study. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4432–4435.

Ercolessi, P., Bredin, H., Sénac, C., and Joly, P. (2011). Segmenting tv series into scenes using speaker diarization. In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 13–15.

Ercolessi, P., Bredin, H., and Sénac, C. (2012a). Stoviz: story visualization of tv series. In *20th ACM international conference on Multimedia*, pages 1329–1330.

Ercolessi, P., Sénac, C., and Bredin, H. (2012b). Toward plot de-interlacing in tv series using scenes clustering. In *10th International Workshop on Content-Based Multimedia Indexing*, pages 1–6.

Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6.

Friedland, G., Gottlieb, L., and Janin, A. (2009). Using artistic markers and speaker identification for narrative-theme navigation of seinfeld episodes. In *11th IEEE International Symposium on Multimedia*, pages 511–516.

Ghaleb, E., Tapaswi, M., Al-Halah, Z., Ekenel, H. K., and Stiefelhagen, R. (2015). Accio: A Data Set for Face Track Retrieval in Movies Across Age. In *ACM International Conference on Multimedia Retrieval*.

Hanjalic, A., Lagendijk, R. L., and Biemond, J. (1999). Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):580–588.

Koprinska, I. and Carrato, S. (2001). Temporal video segmentation: A survey. *Signal processing: Image communication*, 16(5):477–500.

Labatut, V. and Bost, X. (2019). Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5):89.

Li, C. and Chen, G. (2003). Network connection strengths: Another power-law? *arXiv*, cond-mat.dis-nn:0311333.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

McCarthy, P. M. and Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Ratcliff, J. W. and Metzener, D. E. (1988). Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Roy, A., Guinaudeau, C., Bredin, H., and Barras, C. (2014). Tvd: a reproducible and multiply aligned tv series dataset. In *9th International Conference on Language Resources and Evaluation*, page 418–425.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2012). "Knock! Knock! Who is it?" Probabilistic Person Iden-

tification in TV series. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2014a). Story-based Video Retrieval in TV series using Plot Synopses. In *ACM International Conference on Multimedia Retrieval*.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2014b). StoryGraphs: Visualizing Character Interactions as a Timeline. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2015a). Book2Movie: Aligning Video scenes with Book chapters. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2015b). Improved Weak Labels using Contextual Cues for Person Identification in Videos. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Tran, V.-A., Le, V., Barras, C., and Lamel, L. (2011). Comparing multi-stage approaches for cross-show speaker diarization.

Yeung, M., Yeo, B.-L., and Liu, B. (1998). Segmentation of video by clustering and graph analysis. *Computer vision and image understanding*, 71(1):94–109.