How I Met Your Snowclone: Unsupervised Discovery of Snowclone Patterns in Large Datasets

Julien Bezançon[1,2], **Gaël Lejeune**[1], **Marceau Hernandez**[1]
[1]Sorbonne Université, STIH, CERES, 75006, Paris, France
[2]Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
bezancon@lisn.fr

Abstract

Snowclones are a type of Multiword Expression (MWE) pattern that includes open slots, i.e. positions that can be filled with various words. For example, in the phrase "May the X be with you," the slot X can be replaced with virtually any noun. A key feature of snowclones is that the original MWE remains recognizable, carrying its meaning into the new form. However, previous work has not shown whether such substitutions are limited to fixed positions. In practice, variations such as "May the force **bee** with you" are also possible. In this paper, we propose to use Locality Sensitive Hashing (LSH) to automatically extract snowclone patterns from the non-commercial IMDb dataset. This process results in the creation of the FROST lexicon, comprising 30,826 pattern candidates and 1,059,824 snowclone candidates distributed in 30 languages. We then annotate 1,500 discovered patterns and 1,000 snowclones from the FROST lexicon to assess its quality. Our findings suggest that (i) most substitutions in snowclones occur at consistent positions and (ii) snowclones can be reliably discovered at scale using LSH and similarity-based metrics. This work provides the first large-scale lexicon of snowclone-based MWEs and a method that can support future research on MWEs and snowclones discovery.

Keywords: snowclone, multiword expressions, locality sensitive hashing, dataset, lexicon

1. Introduction

Over the years, Multiword Expressions (hereafter MWEs) have shown to be a real "pain in the neck" for NLP (Sag et al., 2002). MWEs correspond to combinations of words with idiomatic features at the lexical, syntactic, semantic, pragmatic, and/or statistical levels (Baldwin and Kim, 2010). They have been shown to be both idiosyncratic and pervasive across different languages (Ramisch, 2023).

One of their main features is their ability to vary from one occurrence to another (Pasquer, 2019), as illustrated in examples (1) and (2) below. We identify two main categories of variation: minor variations, which have limited impact on the status of an MWE, and defixation [1], which can affect both the form and the meaning of the MWE. The latter is widely used to create puns and wordplay based on MWEs, such as example (3), and has been studied to some extent, particularly in the context of French linguistics literature (Fiala and Habert, 1989; Mejri, 2009).

1. "*l'attente en **valait** la peine*"
   (FR, the wait **was** worthwhile)
2. "*l'attente en **vaut** la peine*"
   (FR, the wait **is** worthwhile)
3. "***la tente** en valait la peine*"
   (FR, **the tent** was worthwhile)

The concept of snowclone, closely related to the notion of defixation, was first introduced by Geoffrey K. Pullum, who defined it as "a journalistic cliché phrase with an attention-grabbing hook and totally free parameters for you to set as you wish"[2]. A snowclone corresponds to a pattern based on an existing MWE and allowing for word substitution at some positions (Liberman, 2006). For instance, the MWE "In space, no one can hear you scream" (Alien movie catchphrase, 1979) was used to create the snowclone pattern "In space, no one can hear you X", where X can be replaced by any noun, as in (4) and (5) (Gautier, 2019).

4. "In space, no one can hear you **meow**"
5. "In space, no one can hear you **cry**"

One of the main properties of snowclones, alongside their allowance for word substitution, is that the underlying MWE should remain identifiable, carrying its original meaning into the newly formed variant (Hill, 2018; Traugott et al., 2016). Building on these properties, we argue that it should be possible to (i) automatically discover new snowclones, since their underlying MWE must remain identifiable across occurrences, and (ii) automatically identify the word positions within a snowclone that allow for substitution (hereafter open slots). Most studies on snowclones assume that the underlying template and open slot locations are known in advance. However, none have

[1]A literal translation of "défigement", a term used in French linguistic literature

[2]http://itre.cis.upenn.edu/myl/languagelog/archives/000061.html

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7829–7844
11–16 May 2026.

investigated their unsupervised discovery or verified that substitutions occur *only* at these predefined positions.

In this paper, we use the non-commercial IMDb dataset, composed of 48,209,587 movie and TV show titles distributed in 201 languages, to discover snowclone patterns and their affiliated snowclones. We then construct the FORMULAIC RECOGNITION AND ORGANIZATION OF SNOWCLONE TEMPLATES (FROST) lexicon, which compiles these patterns and their corresponding instances. Our main contributions are as follows:

Automatic snowclone discovery We introduce an unsupervised method using Locality Sensitive Hashing (LSH) to discover snowclone patterns without predefined slots.

FROST lexicon We present the FROST lexicon, the first large-scale resource of snowclone-based MWEs, comprising 30,826 pattern candidates and 1,059,824 snowclone candidates extracted from 30 languages.

Qualitative evaluation We propose a qualitative evaluation across 3 languages, showing that (i) most substitutions in snowclones occur at consistent positions and (ii) snowclones can be reliably discovered at scale using LSH and similarity-based metrics.

Given the scale of the IMDb dataset and the novelty of the task, we were unable to directly compare our approach with existing methods, such as those described in Section 2. Moreover, a quantitative evaluation was not feasible due to the limited availability of snowclone-related resources. To compensate for this, we conducted an in-depth qualitative evaluation through two dedicated annotation tasks. The code used and the annotation files are freely available online under the AGPLv3 license[3].

2. Related Work

MWE discovery. MWE discovery involves detecting and adding multiword expressions to a lexicon. This task is considered challenging for several reasons: (i) it primarily relies on statistical methods, (ii) each detected candidate must be analyzed before being added to a lexicon, making the process more complex than manual addition, and (iii) many of the candidates identified through automated discovery are likely already present in specialized lexicons, particularly for resource-rich languages such as English (Ramisch, 2023). Common MWE discovery tasks imply association measures (Evert, 2005; Church and Hanks, 1989; Pedersen, 1996), word substitution (Pearce, 2001; Zhang et al., 2006; Keller and Lapata, 2003) or even the use of parallel corpora (Tsvetkov and Wintner, 2014; de Caseli et al., 2010). In our case, a snowclone discovery task is particularly useful, as it contributes to what is, to our knowledge, the first lexicon of movie-title-based MWEs and provides insights into the most frequent variations observed for each title.

Snowclone processing. To the best of our knowledge, only two approaches have been used to process snowclones in texts. While Church and Hanks (1989) did not explicitly study snowclones, their work on pattern analysis shares conceptual similarities when they studied a pattern like "save X from Y" using association measures. Hartmann and Ungerer (2023) use regular expressions to query two snowclone patterns ("X be the new Y" and "the mother of all X") over the COCA corpus (Davies, 2010). Sweed and Shahaf (2021) also use regular expressions to query over 20 snowclone patterns on Reddit conversations, efficiently creating a corpus of 3,850 snowclone-sentence pairs.

Both of these approaches (i) correspond to an identification task, since they use snowclone patterns to retrieve similar occurrences in large datasets, and (ii) assume that the flexible positions of each pattern are known in advance, using regular expressions to search for these similar occurrences. While this methodology helped the authors to fetch thousands of snowclones, it might leave out occurrences such as "orange is the **Old** Black" or "the **father** of all bombs". The main difference between our methodology and the other approaches presented here is that we do not assume the open slots of a snowclone in advance. Instead, we identify them by clustering similar titles, based on the hypothesis that titles grouped together are likely derived from the same snowclone pattern.

Locality Sensitive Hashing. We aim to use Locality Sensitive Hashing (LSH) to cluster similar titles at scale with approximate similarity while keeping computing cost reasonable. One of the main benefits of LSH, introduced by Indyk and Motwani (1998) and later refined by Gionis et al. (1999) is its ability to retrieve approximate results rather than exact matches. By hashing and grouping items into buckets, LSH reduces the dimensionality of the dataset.

LSH has already seen applications in various areas of Natural Language Processing (NLP). To our knowledge, Ravichandran et al. (2005) were among the first to apply LSH in an NLP setting,

[3] https://github.com/JulienBez/LSHSnowclone

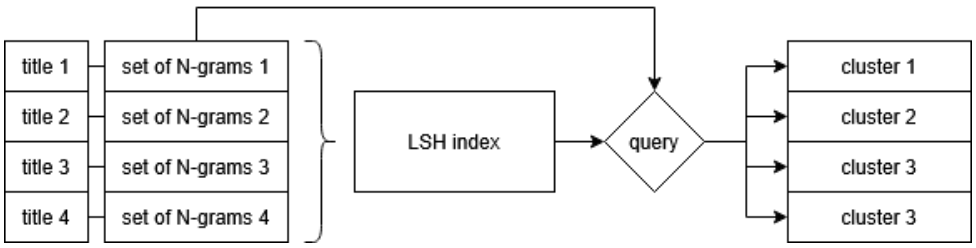

Figure 1: Clustering process using LSH.

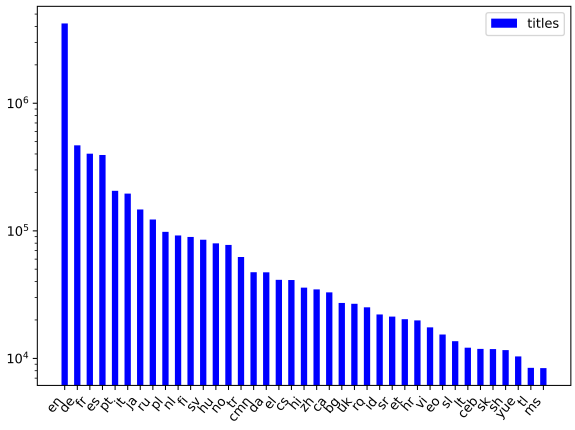

Figure 2: Number of titles for each language.

using it for noun clustering over a dataset of 70 million web pages. Zhao et al. (2020) employed an LSH-based auto-encoder to align the representations of deliberately obfuscated words with their corresponding bullying terms in the context of cyberbullying detection. Similarly, Remil et al. (2024) leveraged LSH to cluster similar bug reports. In this work, we use LSH to create and query indexes of titles.

3. Dataset Description

We use the IMDb non-commercial dataset[4] for our experiments. This dataset is particularly well suited for snowclone discovery, as movie and TV titles are short, often include formulas and catchphrases, and are frequently reused in creative variations. As of March 2026, this dataset contains 48,209,587 movie titles spanning 201 languages, though it includes a considerable amount of noise.

In order to work with this dataset, we performed several processing steps. We (i) removed every malformed row in the dataset, (ii) filtered every title containing the word "episode", as these titles correspond to episode numbers and are uninteresting for our use case, (iii) employed the FASTTEXT-LANGDETECT package (Joulin et al., 2016) to automatically assign a language to each title, compensating for instances where language metadata was missing, and (iv) excluded languages with fewer than 10,000 titles to ensure sufficient data per language for LSH clustering and reliable pattern extraction. As a final step, we took care of duplicates by filtering them out.

In total, 7,301,509 titles and 39 languages remain. Table 11 in the Appendix contains the number of titles found for each language. Figure 2 shows the log scale distribution of titles per remaining language. We observe that more than half of the titles are in English (4,219,258 titles). Table 9 in the Appendix shows the number of titles discarded with each filtering step. In the next Section, we use the remaining titles to create the FROST lexicon.

4. FROST Lexicon Creation

This section describes how we cluster similar titles, recover snowclone patterns, and structure the resulting entries into the FROST lexicon.

4.1. LSH-based Clustering of Titles

The rationale behind our approach is that titles forming a snowclone tend to be lexically similar, making LSH an efficient method for clustering them. The input to this stage of our method is a list of titles, and the output is a cluster of snowclone candidates for each title. For each language, we repeat the following steps: (i) we create an array associating each title with a minhashed set of N-grams of characters composing this title, (ii) we populate the LSH index[5] with every set of N-grams and (iii) for each title in the dictionary, we use its set of N-grams to query the LSH index to find similar titles. Figure 1 illustrates this process. The parameters used for LSH indexation are detailed in the Appendix (A.1). This procedure produces one cluster of similar titles for each input title. In the next step, we recover a snowclone pattern from each cluster and merge clusters with identical patterns.

[4] https://developer.imdb.com/non-commercial-datasets/

[5] We used the DATASKETCH package to hash our titles

(1)		(2)							
le bon, la brute et les zombies		le	bon,	la	brute	et	les	zombies	
le bon, la brute et le truand		le	bon,	la	brute	et	le	truand	
le bon, la brute et le molosse		le	bon,	la	brute	et	le	molosse	
le bon, la belle et le truand		le	bon,	la	belle	et	le	truand	
le bon, la brute et le cinglé	⇒	le	bon,	la	brute	et	le	cinglé	
qui a braqué le boa?		le	bon,	la	brute	et	la	poule	
la brute et la perverse		le	bon,	la	brute	et	le	traitre	
le bon, la brute et la poule					⇓				
le bon, la brute et le traitre		le	bon,	la	brute	et	le	X	(3)

Table 1: Processing of a cluster obtained with the French title "*le bon, la brute et le cinglé*".

4.2. Pattern Recovery

We assume that each cluster of titles contains a reasonable number of snowclones derived from the same MWE. According to the definition of a snowclone, we should be able to identify both fixed and variable components across these titles. For each cluster, we determine the most common title length and retain only titles of that length, as titles with different token counts are often unrelated. Focusing on the dominant length ensures that we compare titles that (i) are likely to be similar and (ii) share a similar structure. This makes comparisons faster and more straightforward. We then align every title to identify the most frequent words at each position within these titles. If a word appears in more than half of the titles at a given position, we treat it as a fixed component of the snowclone pattern. Otherwise, we classify the position as flexible and replace the word with X to indicate a variable slot.

Table 1 illustrates the three-step extraction on an example cluster obtained with the French movie title "*le bon, la brute et le cinglé*". Step (1) displays the isolated cluster with this title, with titles of the same length highlighted in bold. Step (2) illustrates the alignment of these highlighted titles. Step (3) shows the snowclone pattern obtained with this cluster. For alignment, we split each titles according to white spaces.

4.3. Noise Reduction

At this stage, the FROST lexicon contains N clusters, N being the total number of titles processed with our methodology (7,301,509). We argue that the majority of these clusters neither represent nor contain snowclones and therefore constitute noise. To reduce the proportion of noise in the lexicon, we apply four filters, which we further describe in this Section.

Cluster merge. we merge clusters which share the same snowclone pattern and remove duplicates from within each cluster.

thám tử lừng danh conan: X X X X
Thám Tử Lừng Danh Conan: Ảo Thuật Gia Cuối Cùng Của Thế Kỷ
Thám Tử Lừng Danh Conan: Tàu Ngầm Sắt Màu Đen
Thám Tử Lừng Danh Conan: Nàng Dâu Halloween
Thám Tử Lừng Danh Conan: Mê Cung Trong Thành Phố Cổ
Thám Tử Lừng Danh Conan: Sát Thủ Bắn Tỉa Không Tưởng
Thám Tử Lừng Danh Conan: Kẻ Hành Pháp Zero
Thám Tử Lừng Danh Conan: Quả Bom Chọc Trời
Thám Tử Lừng Danh Conan: Thủ Phạm Trong Đôi Mắt
Thám Tử Lừng Danh Conan: 15 Phút Tĩnh Lặng
Thám Tử Lừng Danh Conan: Cơn Ác Mộng Đen Tối

Table 2: Vietnamese titles linked to the snowclone pattern "*thám tử lừng danh conan: X X X X*". We highlight in bold the fixed positions in each title.

Snowclone pattern	Cluster size
X ljubavi	35
X zemlja	34
cirkusrevyen X	33
posljednji X	32
povratak X	29

Table 3: Some snowclone patterns in Croatian, along with the number of snowclone candidates in their affiliated cluster.

Irrelevant patterns. we discard any snowclone pattern in which at least one third of the positions are marked as flexible, in order to avoid inconsistent or uninformative results (such as "X X of X X"). This filter removes irrelevant clusters, such as the ones containing film series, as illustrated in Table 2. It also discards short patterns, consisting of only one or two words, such as those illustrated in Table 3. We argue that these short patterns are also less informative and should not be considered as snowclones: their structure is far more generic, as they contain too few words to exhibit a distinctive pattern. It is then harder to link a snowclone candidate to these snowclone pattern, since there are too few similarities between them to establish a correlation.

Small clusters. we discard clusters with less than k snowclone candidates. To confirm that a pattern is *de facto* a snowclone pattern, we must be able to identify several sequences based on it.

During our creation of the FROST lexicon, we decided to set k to 20. However, this value can hide a certain number of rare snowclones, especially for languages with fewer titles. It is therefore possible to modify k to create another instance of the FROST lexicon.

Duplicates. We observe that some snowclone candidates are present in several clusters. For instance, the candidate "The Beauty and the Beast Mystery" was found in both the "beauty and the X" and the "the good, the bad and the X" cluster. For each duplicate snowclone candidate, we compute the Levenshtein distance (Levenshtein, 1966) between this candidate and each of its associated pattern to determine the most appropriate cluster. This process remove every snowclone candidate duplicate from the FROST lexicon.

Table 10 in the Appendix summarizes the number of discarded titles with each filtering step. Although these filters may be considered overly restrictive, they remove a significant amount of noise, which makes the FROST lexicon more manageable and easier to navigate.

4.4. Lexicon Structuring and Ranking

In total, 30,826 pattern candidates and 1,059,824 snowclone candidates have been extracted from 30 languages. The remaining 9 languages[6] yielded no snowclones after our filtering steps and were therefore excluded from our study. We present detailed statistics for each language as well as some examples for English, French and Spanish in the Appendix (A.4 and A.5).

Among the languages analyzed, English yielded the highest number of patterns (73.02 % of the total number of patterns) and snowclones (75.45 %). This is easily explained by the high number of titles in English, representing 64.25 % of the total number of titles in the IMDb dataset. We observe an average of 1.01 open slots per snowclone pattern for all languages. We use the snowclone candidates and recovered pattern of each cluster to create an entry in the FROST lexicon. We assign a cosine similarity score between each candidate and its pattern, to rank candidates by closeness. Each entry in the lexicon is then structured as follows:

```
snowclone_pattern = {
    "snowclone_candidate_1": 0.8,
    "snowclone_candidate_2": 0.6,
    "snowclone_candidate_3": 0.3,
    ...
}
```

[6]Chinese, Croatian, Serbian, Estonian, Slovenian, Serbo-Croatian, Slovak, Lithuanian, and Malay

the good, the bad and the X	1,822
once upon a time in X	1,200
a fistful of X	532
the good, the bad, and the X	334
the good, the X and the X	162
the good, the bad & the X	67
for a few X more	45
the good, the X and the ugly	32
the X the bad and the X	22
the good, the bad, & the X	20

Table 4: Snowclone patterns corresponding to Sergio Leone movies along with their number of affiliated snowclone.

The parameters used to compute the cosine similarity scores are detailed in the Appendix (A.2). Our next step is to evaluate the quality of the FROST lexicon. To this end, we conduct two annotation tasks, described in the following Section.

5. Assessing Data Quality

In this section, we evaluate the quality of both the extracted snowclone patterns and their affiliated snowclone candidates through qualitative querying and two annotation studies. We also propose an error analysis on some edge cases encountered while navigating the FROST lexicon.

5.1. Qualitative Check on Popular Titles

We first verify that our lexicon retrieves potentially well-known snowclones by querying patterns derived from iconic movie titles. Here, we consider the titles of several films by Sergio Leone: "Once Upon a Time in America", "Once Upon a Time in the West", "A Fistful of Dollars", "For a Few Dollars More", and "The Good, the Bad and the Ugly". These titles are well-known and highly conventionalized, which suggests they are likely to be used for snowclone creation. Table 4 shows the found snowclone patterns correspond to these titles for English. This confirms that the lexicon contains recognizable snowclone families and that our method consistently captures open-slot structures.

Only one cluster pattern was retrieved for both "Once Upon a Time in America" and "Once Upon a Time in the West", which is not surprising since they share the same structure. Furthermore, we observe that some clusters exhibit highly similar patterns and could be merged, such as those derived from the movie title "The Good, the Bad and the Ugly". However, these instances are relatively rare and were therefore not addressed in the current version of the lexicon.

English		French		Spanish	
the making of X	3,652	une histoire de X	625	las aventuras de X	779
the adventures of X	3,424	la guerre des X	580	el secreto de X	478
the art of X	3,039	les aventures de X	397	la historia de X	477
the X part 1	2,945	le X de la mort	343	el regreso de X	469
the story of X	2,903	le retour de X	341	el X de la muerte	457
night of the X	2,646	le secret de X	338	la X de la muerte	436
X in the dark	2,255	le monde de X	319	el sueño de X	373
return of the X	2,128	le temps des X	313	la muerte de X	343
the story of the X	2,093	la fille du X	268	la X del diablo	328
the man in the X	2,070	je suis un X	264	historia de un X	295
X of the dead	2,070	le voyage de X	254	el X del diablo	288
welcome to the X	2,056	le mariage de X	236	el mundo de X	279
revenge of the X	1,902	la bataille de X	232	la venganza de X	261
battle of the X	1,877	les X de l'amour	230	X en la noche	255
X behind the scenes	1,809	emission du X -	230	el hombre de X	250
the X part 2	1,611	le X de l'amour	222	el viaje de X	245
secrets of the X	1,559	la fille de X	222	el hombre de la X	244
king of the X	1,505	la nuit des X	203	la X del amor	222
the legend of X	1,407	les enfants de la X	200	X en la oscuridad	222
beauty and the X	1,393	je suis une X	199	la casa de los X	217

Table 5: Top 20 snowclone patterns with the highest number of affiliated snowclones found in English, French and Spanish.

5.2. Pattern Annotation

To ensure the quality of the snowclone patterns extracted, we survey the results obtained for 3 languages, namely English, French and Spanish, since they present the highest numbers of snowclones. Table 5 contains the 20 snowclone patterns with the highest number of affiliated snowclone for these languages.We observe that some of those patterns are too generic (such as "the X part 1" or "the making of X") or do not correspond to a snowclone ("*emission du X -*" referring to a TV show, where X represents an episode number). Moreover, Figure 3 illustrates the proportion of positions that allow word substitution in the extracted patterns across all languages. We observe that, in most cases, word substitution occurs on the final position of the template.

We then count the number of valid snowclone patterns among the top 500 most productive pattern candidates for English, French and Spanish. To annotate these patterns, we used the following guidelines:

1. The pattern must belong to the target language. If the pattern is in a language other than the one currently under study, it is not counted. While rare, we did encounter such patterns (see Section 5.4).

2. The variable slot X must not stand for a number. Thus, patterns such as "Teenage Mutant Ninja Turtles X", where X denotes the installment number in a movie series, are excluded.

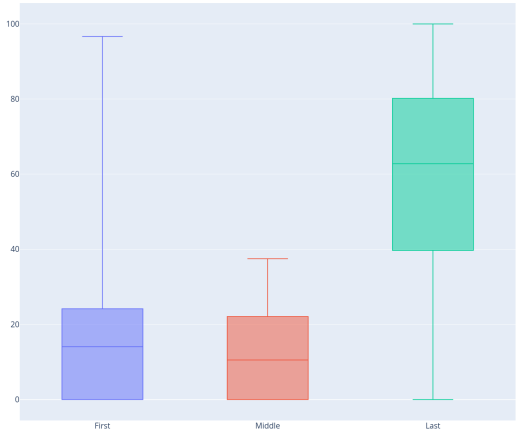

Figure 3: Proportion (in %) of positions allowing for word substitution in the extracted snowclone patterns across all languages.

3. The pattern must not be purely generic. Generic patterns such as "X: A Short Film" or "X A XXX Parody" are excluded, as they provide minimal information about their origin and are often formulaic.

4. At least one known title must match the pattern (without using external knowledge). To validate a snowclone pattern, it must be possible to identify at least one title corresponding to the pattern without relying on external resources or prior knowledge of the affiliated snowclone. For example, the pattern "Interview with a X" can be matched with the title "Interview with a Vampire" (Neil Jordan, 1994).

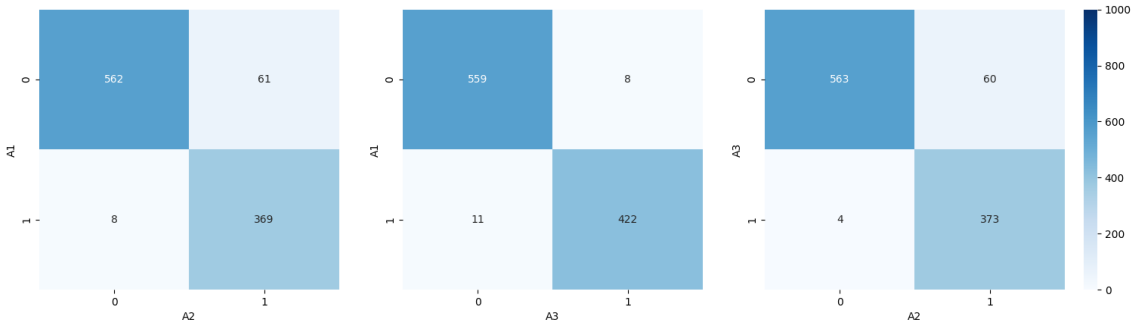

Figure 4: Confusion matrices for each pair of annotator.

Each pattern was annotated by a fluent speaker with high familiarity with snowclone phenomena. We didn't perform cross-annotation, which is why we don't report inter-annotator score for this annotation task. We find 383 (76.60 %) valid snowclone patterns for English, 449 (89.80 %) for French and 440 (88.00 %) for Spanish. These results indicate that the majority of highly productive patterns correspond to genuine snowclones.

5.3. Snowclone Candidates Annotation

In addition to annotating 1,500 snowclone patterns, we evaluate the quality of 1,000 snowclone candidates through a binary annotation task. We perform it on snowclone candidates affiliated to the movie title "The Good, the Bad and the Ugly" (Sergio Leone, 1966). This title was selected for being highly popular and productive, with a total of 1,977 affiliated snowclone candidates distributed in 6 clusters (see Section 5.1). We selected the first 1,000 candidates from the rankings established in Section 4. This second annotation was also conducted by the authors of this paper. Each annotator (hereafter A_1, A_2, and A_3) annotated the same set of candidates to compute inter-annotator agreement. The annotation task was designed as follows: for each candidate, if it corresponds to a valid snowclone, it is assigned the label 1; 0 otherwise. The guidelines provided to the annotators were as follows:

1. Morphological similarity: the candidate should be morphologically related to the pattern or the title that it refers to, containing key common tokens.
2. Phonetic similarity: the candidate should be phonetically related to the pattern or the title that it refers to, having a closely related pronunciation.
3. Semantic similarity: the candidate should be semantically related to the pattern or the title that it refers to, showing an indirect, conceptual connection to them.

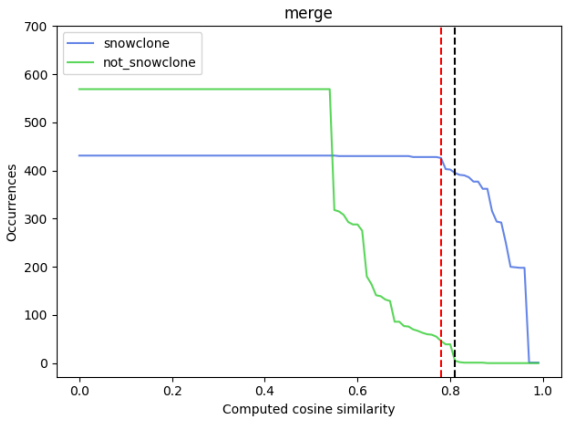

Figure 5: Distribution of candidates according to their assigned label and their cosine score.

The presence of at least one of these criteria was sufficient for a candidate to be considered a true positive. In addition to these guidelines, annotators were provided with the cosine similarity score assigned to each candidate.

We computed a Krippendorff's (Krippendorff, 2013) α score of 0.91, which is notably high for an initial annotation task. Out of 1,000 annotated snowclone candidates, only 65 instances of disagreement were identified. Figure 4 presents the confusion matrices for each pair of annotators. We observe that the annotations from A_2 tend to diverge more frequently from those of A_1 and A_3. To better understand the sources of these discrepancies, an adjudication meeting was conducted during which all three annotators jointly reviewed the cases of disagreement.

During the adjudication meeting, we observed that A_2 considered a fourth criterion: the morphosyntactic category of the substituted words. For example, they annotated the candidate "the good, the bad and the tiger" as not being a snowclone, since "tiger" is a noun whereas "ugly", the original substituted word, is an adjective. During the adjudication, we concluded that a candidate

le temple de shaolin
Le temple de Shaolin
Le Temple de Shaolin
Le Temple de Shaolin 2 : Les Enfants de Shaolin
Le Temple de Shaolin 2: Les Enfants du temple
Le Tigre de Shaolin
5 maîtres de Shaolin
Les jeunes bonzes du temple de Shaolin
Les 18 implacables du temple de shaolin
Les 4 vengeurs de shaolin
Roi de Shaolin
L'aigle de Shaolin
Artes de Shaolin
La fureur des maîtres de shaolin

Table 6: Cluster (in French) for which the snowclone pattern does not contain a flexible position.

remains valid even when the substituted components differ in morphosyntactic category, provided that the underlying pattern is still clearly recognizable. Applying this principle resolved nearly all disagreements, with the remaining ones attributable to minor annotation errors.

In total, we identified 431 snowclone out of 1,000 candidates. Figure 5 shows the distribution of candidates according to their assigned label and score. The red and black dotted lines respectively denote the thresholds at which snowclone identification ceases (0.78) and where true negative candidates begin to appear (0.81). These results, along with the high Krippendorff's α score of 0.91, may indicate that (i) snowclones are relatively easy to identify and (ii) a simple cosine score may effectively distinguish true positives from true negatives.

5.4. Error Analysis

While our methodology allowed us to fetch thousands of snowclone patterns and snowclone candidates, some clusters are less pertinent than others. In this Section, we discuss some edge cases encountered while navigating the FROST lexicon.

Language mismatch. Some titles were attributed the wrong language, leading to the creation of clusters in another language than the one we wanted to study. These errors originate from the FASTTEXT-LANGDETECT package. We suppose that this package, given the average short size of the titles, is prone to making errors when attributing a language. We observe the presence of language mismatch for languages with less than 1,000 snowclone patterns (25 out of 30 languages). In total, 25 clusters in 4 languages correspond to language mismatch, as shown in Table 7. All of these mismatches correspond to English snowclone pattern.

Language	Mismatched clusters
Cebuano (ceb)	18
Hindi (hi)	3
Japanese (ja)	2
Indonesian (id)	1
Swedish (sv)	1

Table 7: Languages for which we observed clusters in another language.

Snowclone pattern	Cluster size
X juni 2013 kl. 12:00	335
X feb. 2016 kl. 12:00	313
X mars 2012 kl. 17:00	296
...	
X feb. 2018 kl. 09:00	13
X mars 2021 kl. 22:55	12
X juni 2022 kl. 19:45	12

Table 8: Snowclone patterns in Norwegian, along with theit number of related snowclone candidates.

No flexible position. Our methodology relies on the snowclone candidates within a cluster to extract a snowclone pattern from this cluster. In some cases, there was no identified flexible position for a cluster, such as for the one illustrated in Table 6. This error occurs because we decided not to lowercase titles before processing them with our methodology, which led to the presence of two or more nearly identical titles in some clusters. It then affects the pattern recovery process (Section 4.2).

Similar clusters. Some clusters exhibit closely related snowclone patterns, such as the ones in Table 4. Several clusters obtained snowclone patterns related to the title "the good, the bad and the ugly". While rare, such case can make it harder to navigate through the FROST lexicon: similar titles may be separated in different clusters.

Finally, we notice that Norwegian exhibits an unusal large number of snowclone patterns and snowclone candidates compared with other languages with a similar number of titles. Upon further investigation, we notice that almost all snowclone patterns for Norwegian correspond to dates, as shown in Table 8.

6. Conclusion

In this paper, we proposed the first unsupervised method to discover snowclone templates at scale. We showed that snowclone variation can be modeled and discovered automatically. We introduced the FROST lexicon, comprised of 30,826 pattern candidates and 1,059,824 snowclone candidates across 30 languages.

Due to the nature of the source data and its associated copyright restrictions, we are unable to release the lexicon directly. Nevertheless, we provide our code and annotation files to enable full reproducibility of our experiments. The complete processing pipeline took approximately 1 hour and 30 minutes to run and around 20 minutes when excluding English data.

Given the scale of the IMDb non-commercial dataset, we used Locality Sensitive Hashing (LSH) to cluster similar titles. In addition to being well-suited for large-scale data, this method relies on approximate matching. This is an advantageous property in the context of snowclone discovery, as snowclones inherently exhibit variation across instances.

We observed that, in most cases, the positions available for word substitution within a snowclone pattern are fixed, which implies that MWE defixation tends to follow a recognizable pattern. This observation aligns with the findings of Blache et al. (2018), who note that the identification of an MWE by an individual typically occurs around its second or third word. This may explain why the final position is frequently used for defixation: it allows the reader or listener to recognize the original MWE before realizing it has been altered, thereby producing a surprise effect. These insights may prove valuable for future work on the discovery and identification of MWEs, particularly by accounting for these flexible positions. Our findings support the idea that defixation is structurally constrained, which opens the door to systematic MWE variation modeling rather than treating MWEs as fully fixed units.

Moreover, we determined that snowclone identification is relatively straightforward to perform for both humans and algorithms. The high Krippendorff's α score of 0.91 obtained between three annotators, along with the clear separation between true positive and true negative candidates based on a cosine score, supports this observation.

Future work will focus on: (i) defining automatically similarity thresholds specific to languages to filter candidates efficiently, (ii) extending evaluation to additional languages from the FROST lexicon and (iii) exploring multilingual clustering to discover cross-language snowclone families.

Limitations

Our main limitation lies in the analysis of the results produced by our methodology. The FROST lexicon comprises 30,826 pattern candidates and 1,059,824 snowclone candidates across 30 languages. These numbers, combined with the lack of previous work on snowclone discovery and the absence of a snowclone discovery benchmark, make it difficult to quantify the performance of our methodology. Conducting an in-depth analysis, even on the top k results, would require at least one or two individuals per language, preferably familiar with the concepts of MWEs and snowclones. While we performed such analysis for 1,500 patterns in 3 languages and 1,000 snowclone in English, we have not yet been able to find enough individuals to analyze our results across additional languages.

We also note that the IMDb non-commercial dataset is frequently updated, which is why any reproduction of our experiments will lead to small changes in the FROST lexicon. Most of the time, these changes will consist in the addition of new titles to the dataset. In this work, we used the IMDb dataset from March 2026.

Ethical Considerations

Some titles in the IMDb dataset are pornographic in nature and may be considered upsetting or offensive to some viewers. Discretion is advised when reviewing the data.

Acknowledgments

This work uses data provided by IMDb. The data was accessed under IMDb's Non-Commercial license for academic research purposes. IMDb does not endorse or sponsor this research. Information courtesy of IMDb (https://www.imdb.com). Used with permission.

7. Bibliographical References

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*, 2 edition. Chapman and Hall/CRC.

Philippe Blache, Stéphane Rauzy, Deirdre Bolger, Chotiga Pattamadilok, and Sophie Dufour. 2018. A Dataset for Studying Idiom Processing with EEG. In *Linguistic and Neuro-Cognitive Resources (LiNCR), LREC 2018 Workshop*, pages 18–22, Miyazaki, Japan.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics* -, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Davies, Mark. 2010. *The Corpus of Contemporary American English as the first reliable monitor corpus of English*.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1):59–77.

Stefan Evert. 2005. The Statistics of Word Cooccur-rences: Word Pairs and Collocations.

Pierre Fiala and Benoît Habert. 1989. La langue de bois en éclat : les défigements dans les titres de presse quotidienne française. *Mots. Les langages du politique*, 21(1):83–99.

Antoine Gautier. 2019. Mèmes et snowclones : entre préfabrication et refabrication. *Cahiers de lexicologie*, 114:225–247.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 518–529, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Stefan Hartmann and Tobias Ungerer. 2023. Attack of the snowclones: A corpus-based analysis of extravagant formulaic patterns. *Journal of Linguistics*, pages 1–36.

Ian E. J. Hill. 2018. Memes, munitions, and collective copia: The durability of the perpetual peace weapons snowclone. *Quarterly Journal of Speech*, 104(4):422–443.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, STOC '98, pages 604–613, New York, NY, USA. Association for Computing Machinery.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Frank Keller and Mirella Lapata. 2003. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Mark Liberman. 2006. The proper treatment of snowclones in ordinary english.

Salah Mejri. 2009. Figement, défigement et traduction. Problématique théorique. page 153.

Caroline Pasquer. 2019. *Garder la trace, mettre de l'ordre et relier les points : modéliser la variation et l'ambiguïté des expressions polylexicales*. Phd thesis, Tours, France.

Darren Pearce. 2001. Synonymy in Collocation Extraction.

Ted Pedersen. 1996. Fishing for Exactness.

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics*. thesis, Aix Marseille Université (AMU).

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 622–629, Ann Arbor, Michigan. Association for Computational Linguistics.

Youcef Remil, Anes Bendimerad, Romain Mathonat, Chedy Raïssi, and Mehdi Kaytoue. 2024. Deeplsh: Deep locality-sensitive hash learning for fast and efficient near-duplicate crash report detection. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg. Springer.

Sweed, Nir and Shahaf, Dafna. 2021. *Catchphrase: Automatic Detection of Cultural References*. Association for Computational Linguistics.

Elizabeth Closs Traugott, Graeme Trousdale, Elizabeth Closs Traugott, and Graeme Trousdale. 2016. *Constructionalization and Constructional Changes*. Oxford Studies in Diachronic and Historical Linguistics. Oxford University Press, Oxford, New York.

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources. *Computational Linguistics*, 40(2):449–468.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated Multiword Expression Prediction for Grammar Engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.

Zehua Zhao, Min Gao, Fengji Luo, Yi Zhang, and Qingyu Xiong. 2020. Lshwe: Improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

A. Appendix

A.1. LSH parameters

We used the LSH implementation from the DATASKETCH package to create our LSH index, which is based on Jaccard similarity. The parameters were as follows:

- $num\_perm = 256$
- $ngram\_range = (3, 3)$
- $threshold = 0.5$

num_perm corresponds to the number of hash functions. We used trigrams of characters to construct the minhash signatures. The number of bands ($b = 42$) and rows per band ($r = 6$) were automatically determined by the package according to num_perm and $threshold$. We processed each title by lowering it before adding it to the LSH index.

A.2. Cosine similarity parameters

We used the CountVectorizer from SCIKIT-LEARN package to compute our cosine similarity scores. The parameters were as follows:

- $ngram\_range = (1, 1)$
- $lowercase = True$
- $stopwords = None$
- $analyzer = word$

A.3. Filtering Steps

Table 9 shows the number of titles discarded with each filtering step performed on the IMDb dataset. Table 10 shows the number of discarded snowclone candidates with each filtering step performed on the FROST lexicon.

Filter	N
Episodes	33,919,699
Languages with less than 10k titles	162,665
Duplicates	6,825,714

Table 9: Number of discarded titles at each filtering step performed on the IMDb dataset. We start with 48,209,587 titles and end with 7,301,509 titles.

Filter	N
Cluster merge	3,361,281
Irrelevant patterns + Small clusters	3,330,296
Duplicates	579,106

Table 10: Number of discarded clusters at each filtering step performed on the FROST lexicon. We start with 7,301,509 clusters and end with 30,826 clusters.

A.4. Statistics

Table 11 presents various statistics on the IMDb dataset, including the proportion of titles in each language, as well as their number of tokens, pattern candidates, and potential snowclones. Figure 6 shows the number of remaining snowclones at different cosine similarity thresholds. Employing a higher threshold can help reduce the number of observed snowclones candidates, making the lexicon more manageable.

A.5. Additional Resources

Tables 12 illustrates 16 annotated snowclone pattern candidates (8 true positive and 8 true negative) for English, French and Spanish.

Tables 13, 14 and 15 show some ranked snowclones for the 5 most productive snowclone patterns in English, French and Spanish. We excluded patterns not referring to an MWE (such as ”the making of X”, which is too generic). For each snowclone, we calculated its cosine similarity score with the snowclone pattern related to it in order to rank all snowclones. We used the SCIKIT-LEARN Python package in order to vectorize each snowclone, taking into account unigrams of words to take word substitutions into account.

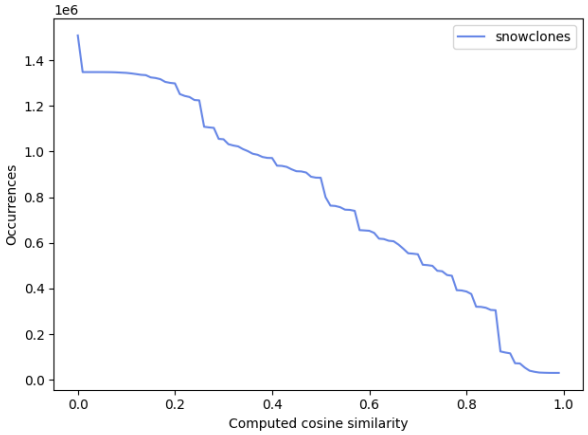

Figure 6: Distribution of snowclones for all languages according to threshold.

	# Title	# Token	Mean	# Pattern	# Snowclone
English(en)	4,219,258	19,182,693	2.11	22,511	799,708
German (de)	467,427	2,009,668	2.43	1,222	38,277
French (fr)	402,273	1,600,558	2.28	2,145	59,041
Spanish (es)	393,205	1,588,540	2.09	2,468	71,026
Portuguese (pt)	205,788	790,107	2.40	1,023	27,216
Italian (it)	195,807	700,390	1.96	586	14,515
Japanese (ja)	147,039	269,096	1.69	3	87
Russian (ru)	122,629	332,249	1.99	20	760
Polish (pl)	98,046	279,359	1.90	10	223
Dutch (nl)	91,916	300,207	1.70	123	2,803
Finnish (fi)	89,489	255,886	1.93	6	148
Swedish (sv)	85,081	286,087	1.86	58	3,061
Hungarian (hu)	79,670	228,683	2.17	17	446
Norwegian (no)	77,477	352,682	2.25	446	37,391
Turkish (tr)	62,286	167,488	1.76	11	446
Mandarin (cmn)	47,349	78,992	1.54	5	67
Danish (da)	47,281	169,865	2.27	17	346
Greek (el)	41,235	142,960	3.13	61	1,392
Czech (cs)	41,107	124,491	2.07	26	733
Hindi (hi)	35,815	93,051	2.42	3	69
Catalan (ca)	32,928	103,494	1.88	13	207
Bulgarian (bg)	27,206	70,828	2.29	5	67
Ukrainian (uk)	26,796	72,000	2.19	5	234
Romanian (ro)	25,122	80,996	2.36	12	418
Indonesian (id)	22,098	62,451	1.49	1	18
Vietnamese (vi)	17,485	72,342	3.34	1	12
Esperento (eo)	15,387	45,124	1.60	7	289
Lithuanian (lt)	12,130	31,812	1.93	1	12
Cebuano (ceb)	11,863	32,215	1.26	18	767
Tagalog (tl)	8,427	35,705	2.35	2	45
Total	7,301,509	29,922,473	2.03	30,826	1,059,824

Table 11: Number of titles, tokens, snowclone patterns candidates and potential snowclones as well as mean token per title in the IMDb dataset (after filtering steps).

Pattern	Label
English	
death of a X	1
a matter of X	1
diary of a X	1
a X to remember	1
return to the X	1
beauty and the X	1
lord of the X	1
X of the living dead	1
X of the century	0
10 things you didn't know about X	0
the case of the X X part 2	0
everything wrong with X X in X minutes or less	0
behind the scenes of X	0
X - part 1	0
the walking dead X	0
world's greatest tv show X	0
French	
le X de la liberté	1
les X de la terreur	1
le silence des X	1
touche pas à mon X	1
bienvenue chez les X	1
la X des étoiles	1
la X des morts-vivants	1
bons baisers de X	1
coup d'oeil nº X	0
analyse et commentaires sur X	0
this ain't X xxx	0
les étrangleurs: X partie	0
les webcolocs... sont X	0
tubes de pub de X	0
X mars 2021 19:45	0
tout le monde X	0
Spanish	
la venganza de los X	1
la X del mal	1
la X de dios	1
el X de los muertos	1
la guerra de la X	1
el triunfo de X	1
guardianes de la X	1
la gran aventura de X	1
festival de cine de san sebastián X - gala de clausura	0
gran premio de X 2024	0
la boca loca de paul - X	0
festival de eurovisión X	0
confetti méxico X de noviembre (2/2)	0
X una historia de amor	0
X luna de miel	0
panorama de actualidad X	0

Table 12: 16 random annotated snowclone patterns for each language, with their label.

Snowclone	Score
le silence des X	
Le silence des églises	0.86
Le silence des ânes	0.86
Le silence des victimes	0.86
Le silence des semences	0.86
Le silence des rizières	0.86
touche pas à mon X	
Touche pas à mon école	0.86
Touche pas à mon témoin	0.86
Touche pas à mon scrab	0.86
Touche pas à mon sport	0.86
Touche pas à mon sponsor	0.86
la guerre des X	
La guerre des étoiles	0.87
La guerre des émeus	0.87
La guerre des écoles	0.87
La guerre des échecs	0.87
La guerre des voisins	0.87
bienvenue chez les X	
Bienvenue chez les fous	0.86
Bienvenue chez les Témakis	0.86
Bienvenue chez les Termites	0.86
Bienvenue chez les Sanders	0.86
Bienvenue chez les SOCCS	0.86
la X des morts-vivants	
La vengeance des morts-vivants	0.89
La survie des morts-vivants	0.89
La secte des morts-vivants	0.89
La révolte des morts-vivants	0.89
La nuit des morts-vivants	0.89

Table 13: Top 5 snowclone for 5 snowclone patterns in French.

Snowclone	Score
el X de la muerte	
El Ángel De La Muerte	0.89
El zombie de la muerte	0.89
El vuelo de la muerte	0.89
El video de la muerte	0.89
El viajero de la muerte	0.89
el secreto de X	
¡El secreto de Violeta!	0.87
¡El secreto de Valentina!	0.87
el Secreto De Mateo	0.87
El secreto de vivir	0.87
El secreto de papá	0.87
el sueño de X	
El sueño de Ícaro	0.87
El sueño de vivir	0.87
El sueño de todos	0.87
El sueño de morfeo	0.87
El sueño de mamá	0.87
las aventuras de X	
Las aventuras de B.J.	1.00
Las aventuras de Walt	0.87
Las aventuras de Ulises	0.87
Las aventuras de Tremendo	0.87
Las aventuras de Tommy	0.87
el hombre de la X	
El hombre de la C.I.A.	1.00
El hombre de la víbora	0.89
El hombre de la tormenta	0.89
El hombre de la sal	0.89
El hombre de la rana	0.89

Table 14: Top 5 snowclone for 5 snowclone patterns in Spanish.

Snowclone	Score
night of the X	
Night of the 1%	1.00
The spirit of the night	0.87
The night of the twins	0.87
The night of the sorcerers	0.87
The night of the mouse	0.87
X of the dead	
The Dead of the Dead	0.96
The will of the dead	0.87
The land of the dead	0.87
The lady of the dead	0.87
The World of the Dead	0.87
the lord of the X	
The lord of the Woods	0.92
The lord of the Wings	0.92
The lord of the Things	0.92
The lord of the Storm	0.92
The lord of the Ruffs	0.92
beauty and the X	
The beauty and the Students	0.87
The beauty and the Sorrow	0.87
The beauty and the Ronin	0.87
The beauty and the Robot	0.87
The beauty and the Nerd	0.87
X of the living dead	
Zombies of the Living Dead	0.89
Year of the Living Dead	0.89
World of the Living Dead	0.89
Weed of the Living Dead	0.89
Warehouse of the Living Dead	0.89

Table 15: Top 5 snowclone for 5 snowclone patterns in English.