The Style of a Successful Story: a Computational Study on the Fanfiction Genre

Andrea Mattei•, Dominique Brunato◇, Felice Dell'Orletta◇
• University of Pisa
a.mattei3@studenti.unipi.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

This paper presents a new corpus for the Italian language representative of the *fanfiction* genre. It comprises about 55k user-generated stories inspired to the original fantasy saga "Harry Potter" and published on a popular website. The corpus is large enough to support data-driven investigations in many directions, from more traditional studies on language variation aimed at characterizing this genre with respect to more traditional ones, to emerging topics in computational social science such as the identification of factors involved in the success of a story. The latter is the focus of the presented case-study, in which a wide set of multi-level linguistic features has been automatically extracted from a subset of the corpus and analysed in order to detect the ones which significantly discriminate successful from unsuccessful stories

1 Introduction

Computational Sociolinguistics is an emergent interdisciplinary field aimed at exploiting computational approaches to study the relationship between language and society (Nguyen et al., 2016). One of the primary factors driving its foundation is the widespread diffusion of social media and other user-generated data available online, which has promoted massive research on computer-mediated communication from several perspectives. For instance, scholars working in the field of genre and register variation have relied on quantitative approaches to inspect the peculiarities of social media language, with the purpose of providing a characterization of this new genre with respect to more traditional ones (Paolillo, 2001; Herring and Androutsopoulos, 2015). In the NLP community, the writing style of user-generated data has been analyzed through computational stylometry approaches for addressing tasks broadly related to author profiling (Daelemans, 2013), such as gender and age detection (Peersman et al., 2011; Koppel et al., 2002). The vast majority of this work has taken into account contents published on few microblogging platforms considered as more representative of the contemporary user-generated mediascape, e.g. Twitter. More recently, the attention has been oriented to the language used by online communities whose members share a common interest towards an object, an activity – and more in general any area of human interest – allowing scholars to shed light on the growing phenomenon of fandom (Sindoni, 2015). One of the most prominent expressions of fandom is *fanfiction* (fanfic, fic or FF), i.e. fiction written by fans of a TV series, movie, book etc., using existing characters and situations to develop new plots. In many languages dedicated websites exist where users can publish their own literary works inspired to the original book they are fans of.

From a computational linguistics standpoint, one perspective from which fanfiction has been investigated aimed to infer the relationship between user-generated stories and their original source, e.g. comparing the representation of characters according to their gender, as well as to model reader reactions to stories (Smitha and Bamman, 2016). Inspired to that study, which was based on a large dataset of stories mainly in English, we collect a new corpus of fanfic stories[1], which, to our knowledge, is the first one for the Italian language. We rely on this corpus to carry out an investigation

[1]Terms of service forbid us to distribute this data. However, the tools used to gather it are available at https://github.com/AndreMatte97/Fanfiction

aimed at shedding light on the possibility of computationally modeling the expected success of a fanfic story, based on the assumptions of linguistic profiling and stylometry research.

2 Dataset collection

The corpus comprises texts collected from efpfanfic.net, a portal active since 2001 which allows users to publish stories and to comment on them. The website is made up of two sections: one for original stories and the other for fanfictions. We considered only the latter and we limited the collection to stories based on the fantasy saga by the British writer J.K. Rowling, "Harry Potter". This choice was motivated by the main purpose of our analysis, i.e. characterizing the success of a novel with respect to its writing style rather than as an effect of the various subject matters it deals with. At the same time, the preference given to a very popular book allowed us to keep a consistent number of potential readers and reviewers across the corpus, still having a large sample of texts to analyze. The data collection was performed through web scraping, with two spiders written in Python using the open-source Scrapy framework. The first spider crawls the list of stories in the category of choice and extracts their first chapters together with some metadata, including the URLs of the subsequent chapters. The second spider takes these addresses as input and downloads texts and additional information about all the chapters after the firsts. In the dataset created this way, the record for each chapter includes: *ID* and *Reference ID*, combinations used by the website to identify the webpage of each chapter. We use the ID of the first chapter as a reference to group together records belonging to the same story; *Title*; *Rating*, an estimate given by the author about the rawness of themes and scenes contained in his story; *Date* of posting; Author's *nickname*; *Number of chapters* in the story; *Text*; Total *number of reviews received by the story*, divided in positive, negative and neutral; *Number of reviews received by the single chapter*, as well as the text of the most recent ones. The crawlers downloaded 54,717 stories, for a total of 19,7310 chapters and a mean of approximately 3.6 chapter per story, which is consistent with the one calculated taking into account every entry on the website. The obtained corpus was divided into folders, each containing stories with the same number of chapters.

3 The success of a fanfiction story: an exploratory study

Based on the newly created dataset, we carried out a computational stylometric analysis aimed at studying whether there is a connection between the success of a fanfic story and its writing style. Such a connection has been demonstrated for more canonical literary works covering novel and movie domains (Ganjigunte et al., 2013; Solorio et al., 2017), showing that stylometry is a viable approach also in scenarios different from authorship attribution and verification.

The methodological framework of our investigation is *linguistic profiling* (Montemagni, 2013; van Halteren, 2004), a NLP-based approach in which a large set of linguistically-motivated features automatically extracted from text are used to obtain a vector-based representation of it. Such representations can be then compared across texts representative of different textual genres and varieties to identify the peculiarities of each. For the purpose of our analysis, we split the original dataset into two varieties corresponding to "successful" and "unsuccessful" stories. To define success we follow an approach similar to that used by Solorio et al. (2017), which is based on the number of reviews obtained by each story. In this regard, we decided to include all reviews, not only the positive ones, which can undoubtedly testify a favorable attitude by the reader for the story. Two main reasons motivated our choice: first, we noticed that the overwhelming majority of collected reviews are written to convey appreciation, with just 0.73% among a total of nearly 900k reviews being negative; therefore, from a statistical point of view, we can reasonably get rid of the distinction between various kinds of reviews and simply take into consideration the overall amount of feedback received. Secondly, also a negative feedback proves that a given story has been read and aroused some interest in the reader. With this in mind, we define as "unsuccessful" those stories that did not receive any reviews, thus being largely ignored by their readers. Conversely, the "successful" category includes all stories with the same number of chapters having received a review count higher than the average of all stories of that length. We also decided to limit the focus of this analysis to single-chapter fanfictions written before 2018, so as to avoid the inclusion of stories not yet concluded. The resulting classes comprise 2101 un-

successful texts and 14486 successful ones, with a threshold for success amounting to 5 reviews. Table 1 shows an example of stories classified in the two categories.

All texts were pre-processed by means of regular expressions, with the aim of removing errors and inconsistencies in the use of punctuation, capitalization and special characters, in order to increase the reliability of automatic linguistic annotation and the process of feature extraction, which were performed using the Profiling-UD tool (Brunato et al., 2020).

In what follows we first provide an overview of the linguistic features used for our statistical analysis and then we discuss the ones that turned out to be more prominent in successful writing.

3.1 Linguistic Features

The set of features is based on the one described in Brunato et al. (2020) and counts more than 150 features, distributed across distinct levels of linguistic annotation and computed according to the Universal Dependencies (UD) annotation framework. These features have be shown to be effective in a variety of different scenarios, all related to modeling the 'form' of a text, rather than the content: e.g., from the assessment of sentence complexity by humans (Brunato et al., 2018) to the identification of the native language of a speaker from his/her productions in a second language (L2) (Cimino et al., 2018). Specifically, they can be grouped into the following main phenomena:

Raw Text Features: Document length computed as the total number of tokens and of sentences ((*#Tokens, #Sentences* in Table 2); average sentence length and token length, calculated in tokens and in characters, respectively (*Sent length, Word length*).

Lexical Richness: Distribution of words and lemmas belonging to the Basic Italian Vocabulary (De Mauro, 2000) (*BIV_Tok, BIV_Types*) and to the internal repertories (i.e. fundamental, high usage and high availability, *BIV_Fund; BIV_High-US; BIV_High-AV*); Type/Token Ratio, a feature of lexical variety computed as the ratio between the number of lexical types and the number of tokens in the first 100 and 200 words of text (*TTR Lemma*); Lexical density.

Morpho-Syntactic Information: Distribution of all grammatical categories, with respect to the Universal part-of-speech tagset (*UPOS_** and the language specific tagset (*XPOS_**); Distribution of verbs according to tense, mood and person, both for main and auxiliar verbs (*aux_*; V_**)).

Verbal Predicate Structure: Average distribution of verbal roots and of verbal heads for sentences (*VerbHead*); features related to the arity of verbs (i.e. average number of dependents for verbal head, distribution of verbs by arity).

Global and Local Parsed Tree: Average depth of the syntactic tree (*MaxDepth*); average depth of embedded complement chains headed by a preposition; average length of dependency links and of the maximum link (*Links Len; Max Link Length*); relative order of the subject and object with respect to the verb;

Syntactic relations: Distribution of typed UD dependency relations (*dep_**);

Use of Subordination: Distribution of main and subordinate clauses (*Main clause, Subord clause*), average length of subordinate chains, distribution of subordinate chains by length.

4 Data Analysis

For each considered feature we calculated the average value and the standard deviation in the two classes. We the assessed whether the variation between mean values is significant using the Wilcoxon rank sum test. We found that 57% (i.e. 126 out of the 219) of features are differently distributed in a significant way between successful and unsuccessful stories. In Table 2 we report an extract of the most interesting ones.

As it can be seen, successful stories are on average longer in terms of number of tokens and sentences (1, 2), although these sentences are generally shorter (3), suggesting that readers appreciate more a plain writing style. However, when lexical factors are considered, the preference is given to texts exhibiting less frequent words, as suggested by the slightly lower distribution of words belonging to the Basic Italian Vocabulary (5,6) and especially to the Fundamental one (7). Inflectional morphology also appears as a domain of variation between the two classes. Successful fanfictions employ quite more often verbs in the second person (15), a feature typical of narrative writing related to direct speech. On the contrary, we observe a higher distribution of third person verb, specifically auxiliaries, both singular (14) and plural (13), in less successful texts, which can hint at a preference for reported speech.

Label	Example (*Italian*)	Example (*English*)
Successful	La città di Edimburgo era sommersa da una cascata d'acqua. Pioveva. Pioveva da giorni e giorni, senza sosta. Il cielo era illuminato di lampi e scosso da tuoni. Le strade erano vuote. Per la prima volta da giorni, allo scoccare della mezzanotte, la pioggia cessò di colpo. Il silenzio piombò sui quartieri che sembrarono improvvisamente più bui. E in quel silenzio penetrante, l'unico rumore che si riusciva a distinguere era un tac-tac-tac leggero e discontinuo. Proveniva da una finestra. La finestra di una lussuosa casa in centro, l'unica luce accesa a quell'ora. Joanne era davanti al computer, fonte di quel tremolio e scriveva. Batteva le dita sulla tastiera per alcuni istanti, poi si fermava, rileggeva, cancellava e riscriveva. Andava avanti così da giorni. I suoi occhi erano stanchi, ma la sua mente lavorava frenetica. Mancava poco[2].	The city of Edimburgh was flooded by a cascade of water. It was raining. It had been raining for days and days, relentlessly. The sky was lit by lightning and shaken by thunder. The streets were empty. For the first time in days, at the stroke of midnight, the rain stopped abruptly. Silence fell upon the districts that suddenly seemed darker. And in that piercing silence, the only noise that could be recognized was a faint and irregular tac-tac-tac. It was coming from a window. The window of a luxurious house in the city centre, the only light still on at that time. Joanne was in front of the computer, source of that trembling and was writing. She tapped her fingers on the keyboard for a few moments, then stopped, reread, deleted and rewrote. She had been going on like this for days. Her eyes were tired, but her mind was working frantically. Almost there.
Unsuccessful	Il cielo era tetro cosparso di nuvole che sembravano volere annunciare un acquazzone, il vento ulula forte facendo sbattere le finestre violentemente, come se volesse gridare, liberarsi da una rabbia repressa. La donna dai lunghi capelli rosso scuro continuava a fissare la devastazione attraverso il vetro che ora si era appannato dal suo stesso respiro. Aveva lo sguardo malinconico non più illuminato da quella dolce espressione che il riso le donava. Una mano le si poggiò sulla spalla e girò pian piano il volto verso la persona amata che con un ritmo lento cominciò ad accarezzarle le gote che assunsero un colorito roseo alla sua pelle pallida. Chiuse gli occhi come per assaporare quel dolce tocco che ora si era spostato nei suoi capelli. "Non guardare più oltre il vetro" Mormorò la voce con una nota di preoccupazione, apparteneva a James, marito di Lily la donna dai lunghi capelli rossi[3].	The sky was bleak strewn with clouds that seemed to want to announce a downpour, the wind howls loudly making the windows slam violently, as if it wanted to scream, to free itself from a suppressed anger. The woman with the long dark red hair kept staring the devastation through the glass that was now clouded by her own breath. Her melancholic gaze was no longer lit up by that sweet look that laughter gave her. A hand rested on her shoulder and slowly turned her face towards the loved one who started slowly caressing her cheeks which took on a rosy tone on her pale skin. She closed her eyes, as if to savor that sweet touch that had now moved into her hair. "Don't look beyond the glass anymore" Whispered the voice with a note of concern, it belonged to James, husband of Lily the woman with long red hair.

Table 1: An extract of a 'successful' story (the most reviewed one) and of an 'unsuccessful' one.

Focusing on the distribution of morpho-syntactic categories, there is a significant difference in the usage of the most common punctuation marks, commas (25) and full stops (26), which are quite more frequent in highly-reviewed fanfictions. These features relate themselves to the previously observed difference in terms of document length, as texts with more sentences necessarily use punctuation marks to divide them. Additionally we can see that balanced marks (24), i.e. parenthesis and quotation marks, occur more in successful texts, strengthening our previous claim about a more frequent presence of direct speech in this class. At syntactic level, dependency relations are slightly shorter in successful texts, both considering the average value of all dependencies (29) and the value of the maximum dependency link (30). In readability assessment studies, longer syntactic dependencies are typically found in complex texts, and the same holds for deeper syntactic trees. Both these features have lower values in highly-reviewed stories, suggest-

[2]The full story can be found at `https://efpfanfic.net/viewstory.php?sid=607026&i=1`

[3]The full story can be found at `https://efpfanfic.net/viewstory.php?sid=27412&i=1`

	Unsucc		Success	
Feature	**Avg**	**(StDev)**	**Avg**	**(StDev)**
Raw Text Features				
1. # Tokens	1401	(1940)	2120	(2718)
2. # Sentences	78.4	(116.7)	125.1	(153.6)
3. Sent length	20.18	(12.39)	17.38	(6.43)
4. Word length	4.50	(.250)	4.52	(.193)
Lexical Features				
5. % BIV_Tok	85.7	(5.1)	84.8	(4.7)
6. % BIV_Types	73.4	(7)	70.1	(7)
7. % BIV_Fund	61	(7.5)	57.1	(7.7)
8. * % BIV_High-AV	3.1	(1)	3.1	(1)
9. % BIV_High-US	8.5	(2.4)	9.1	(2.5)
10. Lexical density	.498	(.033)	.503	(.031)
11. TTR Lemma 100	.560	(.118)	.560	(.112)
12. TTR Lemma 200	.433	(.114)	.436	(.110)
Morpho-Syntactic Features				
13. % Aux_3perPl	13.2	(9)	11.8	(7.6)
14. % Aux_3perSin	54.4	(17.1)	53.2	(15.6)
15. % Aux_2perSin	6.3	(8.1)	7.9	(8.5)
16. % Aux_Imperf.	38.5	(24.8)	31.3	(23.6)
17. % Aux_Pres.	52.4	(26)	60	(23.6)
18. % V_Gerund	5.7	(3.8)	6.3	(4)
19. % upos_VERB	12.5	(1.8)	12.3	(1.7)
20. % upos_NOUN	13.8	(2.3)	13.5	(2.1)
21. % upos_ADJ	4.7	(1.4)	4.6	(1.2)
22. % upos_PRON	8.59	(2.24)	8.51	(2)
23. % upos_ADP	10.8	(1.9)	10.4	(1.8)
24. % xpos_FB	1.7	(2)	2.3	(2.4)
25. % xpos_FF	6.5	(2.7)	7.1	(2.8)
26. % xpos_FS	5.5	(2.2)	6.1	(2.1)
27. % xpos_CC	3.1	(.9)	2.9	(.8)
28. % xpos_CS	1.7	(.7)	1.8	(.7)
Syntactic Features				
29. Links Len	2.78	(.438)	2.72	(.385)
30. *Max Link Len	1.19	(2.38)	.687	(1.33)
30. MaxDepth	3.96	(1.45)	3.58	(.857)
32. % Main clause	48.8	(9.9)	49.9	(9)
33. % Subord clause	51.2	(9.9)	50.1	(9)
34. % Verb Head	2.63	(1.72)	2.26	(.897)
35. % dep_nsubj	4.9	(1.1)	4.7	(1)
36. * % dep_obj	5.3	(1.1)	5.3	(1)
37. % dep_obl	5.5	(1.1)	5.2	(1)
38. % dep_punct	14.2	(3.8)	16	(4.1)
39. % dep_conj	4	(1.3)	3.7	(1.1)
40. % dep_det	10.9	(2)	10.5	(1.8)

Table 2: An extract of linguistic features varying significantly between *successful* and *unsuccessful* stories. All differences are significant at $p < 0.001$, except for features marked with an asterisk, which have $p < 0.05$.

ing that the style of successful writing is characterized by a simpler syntactic structure. Interestingly, these results, although preliminary, go in the opposite direction to those reported by Ganjigunte et al. (2013) for successful literary works in English, which where found to be less correlated with text readability scores. Finally, subordinate clauses (33) occur slightly more often than main clauses (32) in unsuccessful texts, while there is a nearly even split between hypotaxis and parataxis in successful ones.

To deepen our analysis, we also computed the coefficient of variation σ^* for all features varying significantly between the two classes, where σ^* is the ratio between the standard deviation σ and the mean μ. This allowed us to evaluate the dispersion of values around the average in a standardized way, and thus to compare the stability of features pertaining to data measured on different scales. A feature that is much scattered in a class of texts and highly stable in the other has a greater chance of being a meaningful representative of the latter.

In Figure 1 we show the average variability in the two classes of the four groups of features distinguished according to the level of annotation they were extracted from. As a whole, we noticed that successful texts display less variability in nearly every considered feature: 117 of them (92%) are more stable in this class. In successful stories, features with greater stability compared to the other class are mainly raw text, e.g. number of sentences, number of tokens and syntactic ones, e.g. verbal heads per sentence and average depth of syntactic trees. Among the few features which are more stable in poorly received texts, we find instead verbal predicate features, such as the distributions of past tenses and of indicative moods, in addition to the frequency of usage of cardinal numbers. The set of lexical features is instead the most stable one for both classes.

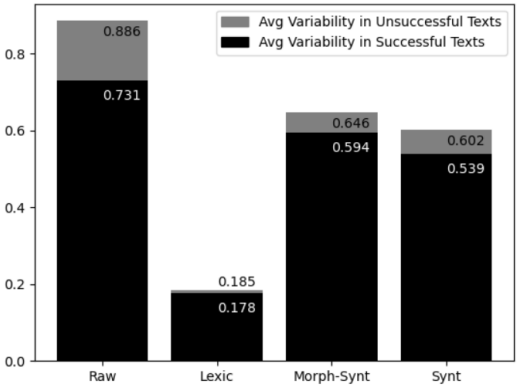

Figure 1: Average coefficient of variation in each class of features, both for successful and unsuccessful texts.

5 Conclusion

In this paper, we presented a NLP-based stylometric analysis on the emerging genre of fanfiction aimed at characterizing the writing style of a successful story. We collected a new large-scale corpus which – to the best of our knowledge – is the first one of this genre for Italian. We showed that successful stories, defined as those receiving a number of reviews higher that the average, are characterized by a variety of linguistic features at different levels of granularity and that these features are more uniformly distributed within them.

In the future, we would like to broad the perspective to other genres in order to study whether there are linguistic predictors of successful writing which are constant across different genres, as well as across concepts somehow similar to success, such as virality and engagement.

References

D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi and S. Montemagni. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, 7145–7151.

D. Brunato, L. De Mattei, F. Dell'Orletta, B. Iavarone and G. Venturi. 2018. Is this Sentence Difficult? Do you Agree? *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, 2018.

A. Cimino, F. Dell'Orletta, D. Brunato and G. Venturi. 2018. Sentences and Documents in Native Language Identification. *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-IT)*, 1–6, Turin.

W. Daelemans. 2013 Explanation in Computational Stylometry. *Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2013*, Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg.

Tullio De Mauro. 2000. *Grande dizionario italiano dell'uso* (GRADIT). Torino, UTET.

V. Ganjigunte Ashok, S. Feng and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1753–1764.

S.C. Herring and J. Androutsopoulos. 2015. Computer-mediated discourse 2.0. *The handbook of discourse, 2nd ed. Deborah Tannen*, Heidi E. Hamilton, Deborah Schiffrin, eds. John Wiley Sons., 1753–1764.

M. Koppel, S. Argamon and A. Rachel Shimoni 2002. Automatically Categorizing Written Texts by Author Gender. *Lit. Linguistic Comput., 17, 4, 401–412*

S. Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, 145–172.

D. Nguyen, A.S. Doğruöz, C.P. Rosé, and F.M.G. de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, Vol. 42, No. 3, 537–593.

John Paolillo. 2001. Language variation on Internet Relay Chat: A social network approach. Journal of Sociolinguistics, 5, 180–213.

C. Peersman, W. Daelemans, and L. Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, 37–44.

M.G. Sindoni 2011. 'I Really Have No Idea What Non-Fandom People Do with Their Lives.' A Multimodal and Corpus-Based Analysis of Fanfiction. *Lingue e Linguaggi, (13), 2015, 277–300, doi.org/10.1285/i22390359v13p277*.

M. Smitha and D. Bamman. 2016. Beyond Canonical Texts: A Computational Analysis of Fanfiction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016.

T. Solorio, M. Montes-y-Gómez, Suraj Maharjan, J. Ovalle and Fabio A. González. 2017. Multi-task Approach to Predict Likability of Books. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1217–1227.

H. van Halteren 2004. Linguistic profiling for author recognition and verification. *Proceedings of the Association for Computational Linguistics*, 200–207.