# Character Identification Refined: A Proposal

**Labiba Jahan & Mark A. Finlayson**
School of Computing and Information Sciences
Florida International University
11200 S.W. 8th Street, CASE 362, Miami, FL 33199
{ljaha002, markaf}@fiu.edu

## Abstract

Characters are a key element of narrative and so character identification plays an important role in automatic narrative understanding. Unfortunately, most prior work that incorporates character identification is not built upon a clear, theoretically grounded concept of character. They either take character identification for granted (e.g., using simple heuristics on referring expressions), or rely on simplified definitions that do not capture important distinctions between characters and other referents in the story. Prior approaches have also been rather complicated, relying, for example, on predefined case bases or ontologies. In this paper we propose a narratologically grounded definition of character for discussion at the workshop, and also demonstrate a preliminary yet straightforward supervised machine learning model with a small set of features that performs well on two corpora. The most important of the two corpora is a set of 46 Russian folktales, on which the model achieves an $F_1$ of 0.81. Error analysis suggests that features relevant to the plot will be necessary for further improvements in performance.

Characters are critical to most of definition of narrative. As an example, Monika Fludernik defines a narrative as "a representation of a possible world . . . at whose centre there are *one or several protagonists* of an anthropomorphic nature . . . who (mostly) perform goal-directed actions . . . " (Fludernik, 2009, p.6; emphasis ours). Therefore, if we wish to advance the field of automatic narrative understanding, we must be able to identify the characters in a story.

Numerous prior approaches have incorporated character identification in one way or another. Some approaches, e.g., examining charaters' social networks (e.g., Sack, 2013), take character identification for granted, implementing heuristic-driven identification approaches over named entities or coreference chains that are not examined for their efficacy. Other approaches have sought to solve the character identification task specifically, but have relied on domain-specific ontologies (e.g., Declerck et al., 2012) or complicated case bases (e.g., Valls-Vargas et al., 2014). Others have taken supervised machine learning approaches (Calix et al., 2013). Regardless, all of the prior work has, unfortunately, had a relatively impoverished view of what a character is, from a narratological point of view. In particular, a key aspect of any character is that it *contributes to the plot*; characters are not just any animate entity in the narrative. We outline this idea first, before describing how we constructed two annotated datasets reflecting this narratologically grounded view oharacter. Then we demonstrate a straightforward supervised machine learning model that performs reasonably well on this data. This paper is just a first proposal on this approach, as much remains be done.

The paper proceeds as follows. First we discuss a definition of character drawn from narratology, contrasting this concept with those used in prior computational work (§1). We then describe our data sources and annotation procedures (§2). Next we discuss the experimental setup including the features and classification model (§3). We present the results and analyze the error patterns of the system, discussing various aspects, which leads us to a discussion of future work (§4). Although we have discussed prior work briefly in the introduction, we summarize work related to this study (§5) before we conclude by enumerating our contributions (§6).

## 1 What is a Character?

All prior works that we have found which incorporate character identification in narrative did not



*Proceedings of the First Workshop on Narrative Understanding*, pages 12–18
Minneapolis, Minnesota, June 7, 2019. 

provide a clear definition of *character*. So far the work that reports the best performance is by Valls-Vargas et al. (2014), where they mentioned different types of characters such as humans, animals (e.g., a talking mouse), anthropomorphic objects (e.g., a magical oven, a talking river), fantastical creatures (e.g., goblins), and characters specific to the folklore (e.g., the Russian characters Morozko and Baba Yaga). Despite this relatively comprehensive list of character examples, they did not provide any properties that distinguish characters from other animate entities.

Consider the follow example. Let's assume we have a story about Mary, a little girl who has a dog named Fido. Mary plays with Fido when she feels lonely. Also, Fido helps Mary in her daily chores and brings letters for Mary from the post office. One day Mary and Fido are walking through town observing the local color. They see a crowd gathered around a fruit vendor; an ugly man crosses the path in front of them; another dog barks at Fido. Many narratologists and lay people would agree that the story has at least two characters, Mary and Fido. Depending on how the story is told, either Mary or Fido may be the protagonist. But what about the other entities mentioned in the story? What about the unnamed man who crosses their path? Is he a character? What about the formless crowd? Is the crowd itself a character, or perhaps its constituent people? What about the fruit vendor, who is hawking his wares? And what about the barking dog? Where do we draw the line?

To clarify these cases, our first goal was to find an appropriate definition of character grounded in narrative theory. We studied different books and literature reviews on narratology that provided different definitions of character. Helpfully, Seymour Chatman, in his classic book "Story and Discourse: Narrative Structure in Fiction and Film" (1986), collected a number of view on character across multiple narratological traditions. Several of the definitions were complex and would be quite difficult to model computationally. Others were too vague to inform computational approaches. However, one definition provided a reasonable target:

> The view of the Formalists and (some) structuralists resemble Aristotle's in a striking way. They too ague that characters are products of plots, that their status is "functional," that they are, in short, participants or *actants* rather than *personnages*, that it is erroneous to consider them as real beings. Narrative theory, they say, must avoid psychological essences; aspects of character can only be "functions." They wish to analyze only what characters do in a story, not what they are—that is, "are" by some outside psychological or moral measure. Further, they maintain that the "spheres of action" in which a character moves are "comparatively small in number, typical and classable." (Chatman, 1986, p.111)

Here, an *actant* is something that plays any of a set of active roles in a narrative and *plot* denotes the main events of a story. This definition, then, though presented via somewhat obscure narratological terminology, gives a fairly conceptually concise definition of a character: A character is *an animate being that is important to the plot*. By this measure then, we are justified in identifying Mary and Fido as characters, but not the various entities they casually encounter in their stroll through town.

## 2 Data

Armed with this refined definition of character, we proceeded to generate preliminary data that could be used to explore this idea and demonstrate the feasibility of training a supervised machine learning system for this concept of character. We sought to explore how easily computable features, like those used in prior work, could capture this slightly refined concept of character. We began with the fact that characters and other entities are expressed in texts as coreference chains made up of referring expressions (Jurafsky and Martin, 2007). Thus any labeling of *character* must apply to coreference chains. We generated character annotations on two corpora, one with 46 texts (the extended ProppLearning corpus) and other with 94 texts (a subset of the InScript corpus), for a total of 1,147 characters and 127,680 words.

The ProppLearner corpus was constructed for other work on learning plot functions (Finlayson, 2017). The corpus that was reported in that paper comprised only 15 Russian folktales, but we obtained the extended set of 46 tales from the authors. These tales were originally collected in

| | Texts | Tokens | Coreference Chains | | | | |
|---|---|---|---|---|---|---|---|
| | | | Total | Anim. | Inanim. | Char. | Non-Char. |
| ProppLearner (Ext.) | 46 | 109,120 | 4,950 | 2,004 | 2,946 | 1,047 | 1,361 |
| Inscript (Subset) | 94 | 18,568 | 615 | 105 | 510 | 94 | 521 |
| Total | 140 | 127,680 | 5,565 | 2,098 | 3,467 | 1,141 | 1,882 |

Table 1: Counts across coreference chains of different categories, as well as texts and tokens.

| Coreference Chain Head | Class | Explanation |
|---|---|---|
| Nikita, tsar | Character | People who perform as a character |
| he, she, her | Character | Animate pronouns that perform as a character |
| walking stove, talking tree | Character | Inanimate entities that perform as a character |
| a bird, insects | Non Character | Animate entities that does not perform as a character |

Table 2: Examples of annotation of characters in coreference chain level.

Russian in the late 1800's but translated into English within the past 70 years. All of the texts in the corpus already had gold-standard annotations for major characters, congruent with our proposed definition. Usefully, the corpus also has gold-standard annotations for referring expressions, coreference chains, and animacy.

We also investigated the InScript corpus (Modi et al., 2017). InScript contains 1,000 stories comprising approximately 200,000 words, where each story describes some stereotypical human activity such as going to a restaurant or visiting a doctor. We selected a subset (94 stories, approximately 19k tokens) of the corpus that describes activity of taking a bath. It has referring expressions and coreference chains already annotated.

The first author manually annotated both of these corpora as to whether each coreference chain acted as a character in the story. According to the definition mentioned above, we marked a chain as character if it is animate and participates in the plot of the story. Because this is a preliminary study, we have not yet done double annotation; this will done as be future work. According to our definition, characters must be animate; thus, because the ProppLearner corpus provides gold-standard animacy markings, on that corpus we only assessed whether animate chains represented characters. The InScript corpus did not come with animacy markings, and so we assessed every coreference chain. The stories in the InScript corpus are fairly simple, and usually only involve a single protagonist, alone in the story. Because of this, every single animate chain in that data was also a character, and both automatic animacy detection and character detection worked extremely well; as we will discuss later, this is rather uniformative. Table 1 shows the total number of texts and tokens in each corpus, as well as a breakdown of various categories of coreference chain: animate, inanimate, character, and non-character. Table 2 gives some examples of character annotations.

## 3 Approach

Because to be a character a referent must actively involved in the plot, characters are necessarily animate, although clearly not all animate things are necessarily characters. Animacy is the characteristic of independently carrying out actions in the story world (e.g., movement or communication) (Jahan et al., 2018). Therefore detecting the animacy of coreference chains will immediately narrow the set of possibilities for character detection. Our character identification system thus consists of two stages: first, we detect animate chains from the coreference chains using an existing animacy detector (§3.1); second, we apply a supervised machine learning model that identifies which of those chains qualify as characters (§3.2).

### 3.1 Animate Chain Detection

Our first step was to identify animate chains. In order to do that we used a coreference animacy detector described in prior work (Jahan et al., 2018). This model is a hybrid system incorporating both supervised machine learning and hand-built rules, and achieves state-of-the-art performance. The extended ProppLearner corpus came with animacy

| Corpus | Acc. | Inanimate | | | | Animate | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\kappa$ | Prec. | Rec. | $\mathbf{F_1}$ | $\kappa$ | Prec. | Rec. | $\mathbf{F_1}$ |
| ProppLearner | 85% | 0.72 | 0.93 | 0.82 | 0.87 | 0.72 | 0.78 | 0.92 | 0.84 |
| InScript | 99% | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

Table 3: Performance of the animacy model on the corpora.

| Corpus | Feature Set | Acc. | Non Character | | | | Character | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\kappa$ | Prec. | Rec. | $\mathbf{F_1}$ | $\kappa$ | Prec. | Rec. | $\mathbf{F_1}$ |
| Propp-Learner | Baseline MFC | 56% | 0.0 | 0.57 | 1.0 | 0.72 | 0.0 | 0.0 | 0.0 | 0.0 |
| | SS, WN, NE | 80% | 0.82 | 1.0 | 0.87 | 0.93 | 0.64 | 0.75 | 0.80 | 0.77 |
| | WN, CL | 80% | 0.82 | 1.0 | 0.87 | 0.92 | 0.64 | 0.75 | 0.80 | 0.78 |
| | CL, SS, WN | 84% | 0.78 | 1.0 | 0.84 | 0.92 | 0.66 | 0.75 | 0.84 | 0.79 |
| | CL, WN, NE | 82% | 0.81 | 0.86 | 0.92 | 0.92 | 0.64 | 0.82 | 0.77 | 0.80 |
| | CL, SS, WN | 84% | 0.78 | 1.0 | 0.84 | 0.92 | 0.66 | 0.75 | 0.84 | 0.79 |
| | CL, SS, WN, NE | 85% | 0.78 | 1.0 | 0.85 | 0.91 | 0.66 | 0.88 | 0.76 | **0.81** |
| InScript | CL, SS, WN, NE | 99% | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | **0.99** |

Table 4: Performance of different features sets for identifying characters. MFC = most frequent class. $\kappa$ = Cohen's kappa (Cohen, 1960)

already marked; the InScript corpus already has gold standard coreference chains, and so we used those coreference annotations as input to the animacy model to generate animacy markings. The performance of the animacy model on both corpora is shown in Table 3.

### 3.2 Feature Selection for Character Identification

We used four different features for our character identification model.

1. **Coreference Chain Length (CL)**: We computed the length of a coreference chain as an integer feature. This feature explicitly captures the tendency of the long chains to be characters, as discussed in prior work (Eisenberg and Finlayson, 2017).

2. **Semantic Subject (SS)**: We also computed whether or not the head of a coreference chain appeared as a semantic subject (ARG0) to a verb, and encoded this as a boolean feature. We used the semantic role labeler associated with the Story Workbench annotation tool (Finlayson, 2008, 2011) to compute semantic roles for all the verbs in the stories.

3. **Named Entity (NE)**: We computed whether or not the head of a coreference chain appeared was a named entity with the category *PERSON*, and encoded this as a boolean feature. The named entities were computed using the classic API of the Stanford dependency parse (Manning et al., 2014, v3.7.0).

4. **WordNet (WN)**: We checked if the head of a coreference chain is a descendant of *person* in WordNet, and encoded this as a boolean feature.

### 3.3 Classification Model

Our classification model is straightforward supervised machine learning, in which we explored different combinations of our features. We implemented our model using an SVM (Chang and Lin, 2011) with a Radial Basis Function Kernel[1]. We tested different combinations of features on the ProppLearner corpus, and their relative performances are shown in Table 4. The best performing feature set was using all four features, and we also tested this model on the InScript data. We trained each model using ten-fold cross validation, and report macroaverages across the performance on the test folds.

## 4 Results, Error Analysis, & Discussion

The best model, using all four features, achieves an $F_1$ of 0.81 on the ProppLearner data, and an $F_1$ of 0.99 on the InScript data. The result on the InScript data is misleadingly high and deserves some

[1]SVM parameters were set at $\gamma = 1$, $C = 0.5$ and $p = 1$.

discussion. The InScript stories are quite simple, only told in the first person, and usually featuring only a single animate referent who is also the protagonist. Therefore the almost exclusive reference to characters in these stories was the personal pronoun *I*. Thus both the animacy detector and the character identifier had much higher performance than one would expect on more complicated stories.

A detailed error analysis of the results on the ProppLearner data revealed at least three major problems for the character identification model.

**First**, the character model relied on the output of the animacy model, and so if a character was not marked animate, the character model also missed it. Conversely, sometimes inanimate chains are incorrectly marked animate, providing an additional opportunity for the character model to err. Thus, in order to improve the performance of our character model, we have to improve the performance of the animacy model.

**Second**, it is hard to detect a character chain with a very few mentions. To solve this problem we could possibly add some new features related to events of the story because event patterns can be helpful to find a character.

**Third**, some non-character animate entities were incorrectly identified as characters, because there is strong correlation between animacy and character. To solve this problem we need more analysis of the plot structure and to find features that more specific to character vis-a-vis animacy.

The last point is critical. Although it seems that features related to how animate and prevalent a referent is are quite useful for identifying characters, they still fall somewhat short. We hypothesize that features related to encoding aspects of the plot, to determine if a referent is contributing to the plot in a meaningful way, will be critical to substantially improving character identification performance. We plan to explore this idea in future work.

## 5 Related Work

The most relevant prior work is a case based reasoning (CBR) system called *Voz* (Valls-Vargas et al., 2014). Voz could identify characters in unannotated narrative text and achieved an accuracy of 93.5%. The system relied on 193 different features. They also proposed a new similarity measure called *Continuous Jaccard* to measure the similarity between a given entity and those in the case base. Although quite useful, this system does not give a theoretically grounded definition of character, and the CBR system is quite complicated.

Calix et al. (2013) developed a model to detect sentient actors in spoken stories. This is akin to animacy detection. They implemented a SVM classifier using 4 categories of features: syntactic, knowledge-based, relation to pronouns, and general context based. Their model achieved 0.86 $F_1$ score, but, because they are focusing on animacy, they are only detecting a set of entities that contain the characters, not the characters themselves.

Declerck et al. (2012) used an ontology-based method to detect characters in folktales. Their ontology consists of family relations as well as elements of folktales such as supernatural entities. After looking at the heads of noun phrases and comparing them with labels in the ontology, they added the noun phrase to the ontology as a potential character if a match was found. Then, they applied inference rules to the candidate characters in order to find two strings in the text that refer to the same character. They discarded strings that were related only once to a potential character and were not involved in an action. They obtained an accuracy of 79%, a precision of 0.88, a recall 0.73, and an $F_1$ of 0.80. Their implicit definition character is most similar to ours, but their ontology based approach is domain specific. As with most domain specific approaches, it would likely not generalize easily to other domains.

Goh et al. (2012) implemented a rule-based system using verbs and WordNet in order to determine the protagonists in fairy tales (where protagonists must by necessity be animate). This is a related task, but not exactly the same as full character identification. They used the Stanford parser's phrase structure trees to obtain the subjects and objects of the verbs and used the dependency structure to obtain the head noun of compound phrases. Additionally, they used WordNet's *derivationally related* relation to find verbs associated with a particular nominal action. They achieved a precision of 0.69, a recall of 0.75, and an $F_1$ of 0.67.

Mamede and Chaleira (2004) developed a system to identify which entities were responsible for the direct and indirect discourses found in children stories. Again, this is a related task but not the

same as character identification. They achieved an accuracy of 84.8% on the training corpus, and 65.7% on the test corpus. Similarly, Zhang et al. (2003) developed a system to identify speakers of the children's story for speech synthesis. In this system they automatically identified quoted texts and assigned speaker to each quote. They did not report the exact performance of their system.

Bamman et al. (2014) developed a hierarchical Bayesian approach to infer latent character types automatically in a collection of 15,099 English novels published between 1700 and 1899. First, they implemented character clustering and then generated related texts to a character to decide which persona a particular character embodies.

Vala et al. (2015) implemented an eight stage pipeline incorporating NER, coreference chains, a series of name variation rules, and WordNet senses to identify characters in literary texts, achieving an an $F_1$ of 0.76.

## 6 Contribution

This paper makes three contributions. First, we proposed a more appropriate definition for *character* in narrative, in contrast to prior computational works which did not provide a theoretically grounded definition.

Second, we singly annotated 46 Russian folktales and 94 InScript stories for character. The InScript stories are unfortunately not as interesting because they contained only a single protagonist each, only ever referred to in the first person.

Finally, we have demonstrated a supervised machine learning classifier for identifying characters, achieving performance of 0.81 $F_1$, which shows that the task is feasible but allows for further improvement.

## Acknowledgments

This work was supported by NSF CAREER Award IIS-1749917. We would also like to thank the members of the FIU Cognac Lab for their discussions and assistance.

## References

David Bamman, Ted Underwood, and Noah A Smith. 2014. A Bayesian mixed effects model of literary character. In *the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 370–379, Baltimore, MD.

Ricardo A Calix, Leili Javadpout, Mehdi Khazaeli, and Gerald M Knapp. 2013. Automatic detection of nominal entities in speech for enriched content search. In *Proceeedings of the 26th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 190–195, St. Pete Beach, FL.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Seymour Chatman. 1986. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, NY.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. Ontology-based incremental annotation of characters in folktales. In *the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 30–34, Avignon, France.

Joshua Eisenberg and Mark Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2708–2715, Copenhagen, Denmark.

Mark A. Finlayson. 2008. Collecting semantics in the wild: The story workbench. In *the AAAI Fall Symposium on Naturally Inspired Artificial Intelligence*, pages 46–53. Arlington, VA.

Mark A. Finlayson. 2011. The Story Workbench: An extensible semi-automatic text annotation tool. In *the 4th Workshop on Intelligent Narrative Technologies (INT4)*, pages 21–24. Stanford, CA.

Mark A. Finlayson. 2017. ProppLearner: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory. *Digital Scholarship in the Humanities*, 32(2):284–300.

Monika Fludernik. 2009. *An Introduction to Narratology*. Routledge, New York.

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. 2012. Automatic identification of protagonist in fairy tales using verbs. In *the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 395–406, Kuala Lumpur, Malaysia.

Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. 2018. A new approach to animacy detection. In *the 27th International Conference on Computational Linguistics (COLING)*, pages 1–12, Santa Fe, NM.

Daniel Jurafsky and James H. Martin. 2007. *Speech and Language Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Nuno Mamede and Pedro Chaleira. 2004. Character identification in children stories. In *The 4th International Conference on Natural Language Processing in Spain (EsTAL)*, pages 82–90, Alicante, Spain.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60. Baltimore, MD.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2017. Inscript: Narrative texts annotated with script information. *arXiv preprint arXiv:1703.05260*.

Graham Alexander Sack. 2013. Character Networks for Narrative Generation: Structural Balance Theory and the Emergence of Proto-Narratives. In *the 4th Workshop on Computational Models of Narrative (CMN'13)*, pages 183–197, Hamburg, Germany.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 769–774, Lisbon, Portugal.

Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. In *the 7th Intelligent Narrative Technologies Workshop (INT7)*, pages 38–44, Milwaukee, WI.

Jason Y Zhang, Alan W Black, and Richard Sproat. 2003. Identifying speakers in children's stories for speech synthesis. In *the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2041–2044, Geneva, Switzerland.