# The Magical Word Vectors of Harry Potter and Tanya Grotter

*Quinn Dombrowski | qad@stanford.edu | @quinnanya*

*https://github.com/quinnanya/dlcl204*

The Harry Potter series is a global phenomenon, having been translated into over 70 languages. Two years after Harry Potter was officially translated into Russian, Dmitry Yemets released the first book in the Tanya Grotter series. While the first book in the series mirrors tropes found in Harry Potter and the Chamber of Secrets (a mistreated orphan living with unpleasant relatives is whisked away to a school for magicians), the choices to use a female protagonist and antagonist, locate the story in Russia, and draw secondary characters from Slavic and Greek mythology result in a distinctly different story.

Time Warner sought to obtain a cease and desist order in the Netherlands, where the first translation of Tanya Grotter was to be published, after being rebuffed in Russian courts. Yemets and his publisher argued that Tanya Grotter was a parody, a protected class of derivative works, and that Harry Potter itself drew heavily on folklore. Nonetheless, Time Warner won the case, preventing the official translation and distribution of the Tanya Grotter series outside of Russia.

Can computational text analysis bring a new perspective to the question of how to quantify the similarity between the magical worlds of Harry Potter and Tanya Grotter? This project uses word vectors as way of comparing these fictional worlds.

## About Word Vectors

The word embedding model popularized by Google's word2vec algorithm has seen increasing adoption for computational literary analysis, by allowing us to explore and analyze networks of conceptual relationships. This project uses the gensim Python package to run word2vec on all 7 Harry Potter novels (as one corpus), all 14 Tanya Grotter novels (as a second corpus). Each word is rendered as a 200-element vector that is meant to capture its semantic profile.

Querying the model for the most similar, and most dissimilar, words to a protagonist's name, is illustrative here. (Note: character names are in green):

**Most similar to Таня 'Tanya'**
Ванька 'Vanka'
она 'she'
Ягун 'Yagun'
Глеб 'Gleb'
он 'he'
Дядя Герман 'Uncle German'
что-то 'something'
Бейбарс 'Baibars'
Пипа 'Pipa'
И-Ван 'I-Van'

**Least similar to Таня 'Tanya'**
хнык-хнык 'hnyk-hnyk'
синица 'titmouse'
нешуточный 'serious'
вменяемый 'sane'
бульонова 'broth'
а-а-а-а-а-а 'a-a-a-a-a-a'
час_спустя 'an hour later'
захлопотать 'shake up'
глава_2 'chapter 2'
* _ *

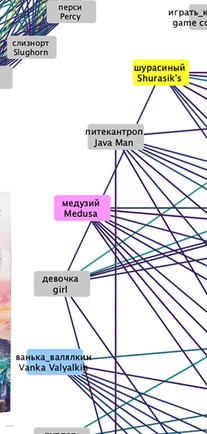





## Character Semantic Similarity

Harry Potter and Tanya Grotter share numerous character tropes: the plucky orphan (Harry/Tanya), the pettily nasty uncle (Vernon/German), the coddled cousin (Dudley/Pipa), the best friend with cultural proficiency in the magical world (Ron/Bab-Yagun), the academically-gifted (potential) romantic interest (Hermione/Vanka), the dark wizard in training (Draco/Coffinia), the senior male wizard whose evil appearance belies a more complicated backstory (Snape/Slander), the stern senior female wizard (McGonagall/Medusa), the aged head of the school (Dumbledore/Sardanapalus). Do these characters share the same semantic relationships amongst themselves?

This network shows a link between major characters when one character appears in the other's list of the 30 most similar words. The color of the link indicates the similarity score, with dark purple showing very similar words, and yellow showing less similar words.



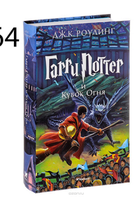



## Similar Concepts, Different Similar Terms

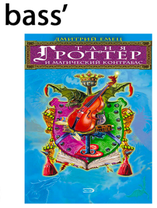
Word vectors trained on very large corpora can be used to produce meaningful analogies (e.g. king : queen :: man : **woman**). Even doing word vectors on both book series combined (34,965 unique items in the vocabulary) isn't enough text to enable analogies across series. For гарри: квиддич :: таня: ?? ('Harry : Quidditch :: Tanya : ??'), where we would expect драконбол 'dragonball', we get characters and words exclusively from the Harry Potter series. Using Tanya as the prompt, we only get characters and words from the Tanya Grotter series. A future expansion of the project to include corpora of Russian Harry Potter and Tanya Grotter fanfic might help improve the vectors' analogy performance. As an alternative, we can look at the most similar words for key parallel terms in each series (vectors trained only on that series).

## Magical conveyances

**метла 'broom'**
кубок 'cup', 0.973
хвост 'tail', 0.970
открытый_небо 'open sky', 0.965
могила 'grave', 0.964
отплывать 'sail away', 0.964
палатка 'tent', 0.964
прицел 'aim', 0.960
зеркало 'mirror', 0.958
одежда 'clothes', 0.958
перебегать 'run across', 0.957



**контрабас 'double bass'**
нога 'leg', 0.961
рука 'arm', 0.953
кольцо 'ring', 0.952
лицо 'face', 0.951
смычок 'bow', 0.950
спина 'back', 0.950
футляр 'case', 0.949
дверь 'door', 0.948
глаз 'eye', 0.948
палец 'finger', 0.947



## Magical places

**Хогвартс 'Hogwarts'**
магл 'Muggle', 0.983
любой 'any', 0.979
год 'year', 0.970
ночь 'night', 0.967
мир 'world', 0.967
заклинание 'spell', 0.964
случай 'occurrence', 0.962
день 'day', 0.960
дементор 'Dementor', 0.959
тот_кто 'the one who', 0.954

**Тибидохс 'Tibidox'**
ученик 'student', 0.965
игрок 'player', 0.956
новый 'new', 0.941
первый 'first', 0.938
любой 'any', 0.937
пять 'five', 0.936
каждый 'each', 0.936
старый 'old', 0.933
десять 'ten', 0.931
мир 'world', 0.931