

Векторные модели художественной литературы: перспективы исследовательского применения



Борис Орехов (НИУ ВШЭ, ИРЛИ РАН)

Немного о векторах

Статья Томаша Миколова, Джефа Дина и Ко (2013)



Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov

Google Inc., Mountain View, CA
tmikolov@google.com

Greg Corrado

Google Inc., Mountain View, CA
gcorrado@google.com

Kai Chen

Google Inc., Mountain View, CA
kaichen@google.com

Jeffrey Dean

Google Inc., Mountain View, CA
jeff@google.com



Abstract

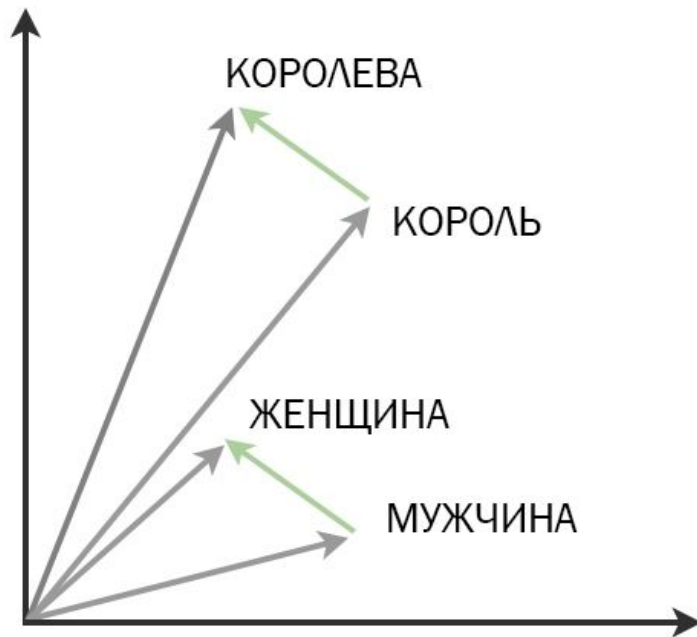
We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

**Векторные модели сделали
лексическую семантику
операционализируемой для
компьютера**

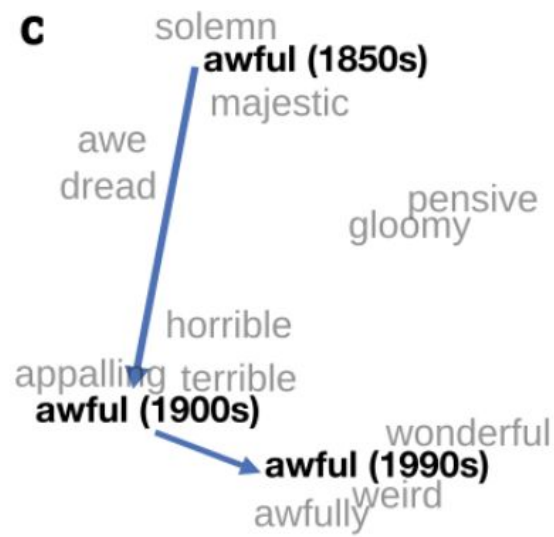
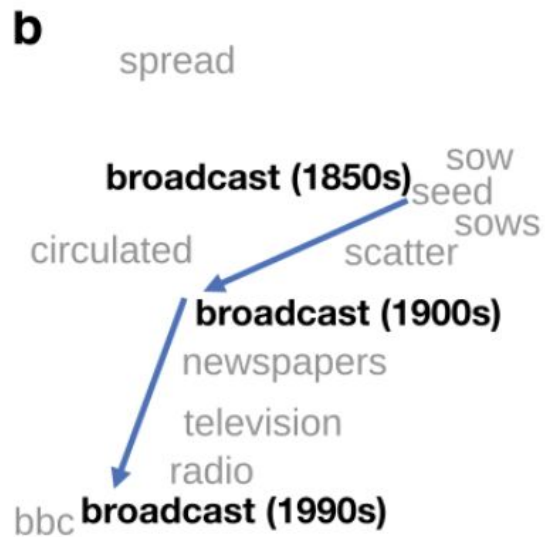
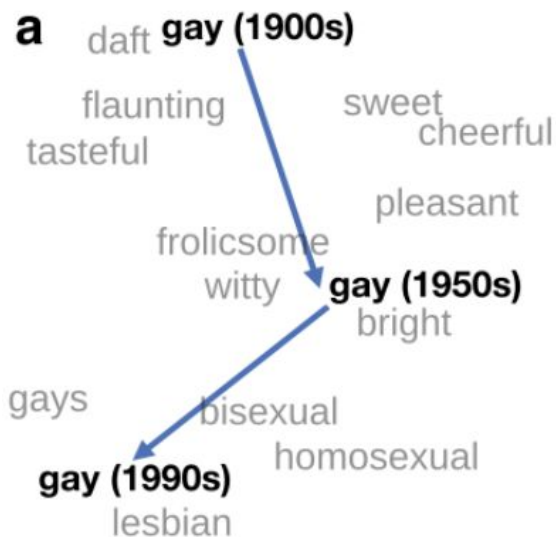




Классический пример семантической математики word2vec



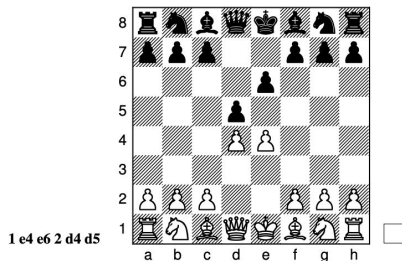
Детектирование семантических сдвигов



Работает даже с шахматными партиями:

B. Orekhov, 'You shall know a piece by the company it keeps. Chess plays as a data for word2vec models'
2024. <https://arxiv.org/abs/2407.19600>

Let's try to do this with moves. We'll take typical opening moves for white and add a move from the same opening for black. King's Gambit: Pe2e4 - Pf2f4 + pe7e5. The model suggests the move pe7e6. This is also an opening move that appears, for example, in the French Defense: Pe2e4, pe7e6, Pd2d4, pd7d5...



Now let's work with the model that includes not only moves but the position on the board (Type 2 model): positions_moves_pro.model. It includes 4946 words. The closest quasi-synonym neighbors are:

->Pe2e4, Δe2e4

->Pe2e3 0.4605116844177246
->Pe3e4 0.3592767119407654
->Bf1g2 0.27666473388671875
->Pe2f3 0.24844536185264587
->Pd2d4 0.2099190503358841

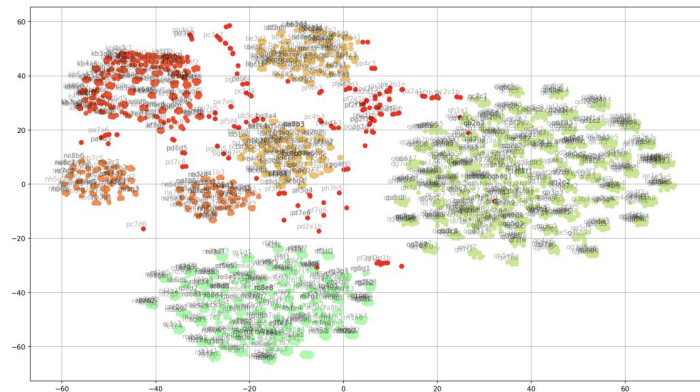



Figure 6: tSNE visualisation of the moves from the endgame_moves_black.model with perplexity 30

In fig. 6, the white-squared bishop is in a different cluster from the black-squared bishop, which makes sense from a game-play perspective.

Векторные модели и художественная литература



Векторные романы

<https://nevmenandr.github.io/novel2vec>



Эмбеддинги персонажей Джейн Остин:

S. Grayson, M. Mulvany, K. Wade, G. Meaney, and D. Greene, Novel2Vec: Characterising 19th Century Fiction via Word Embeddings. 2016.

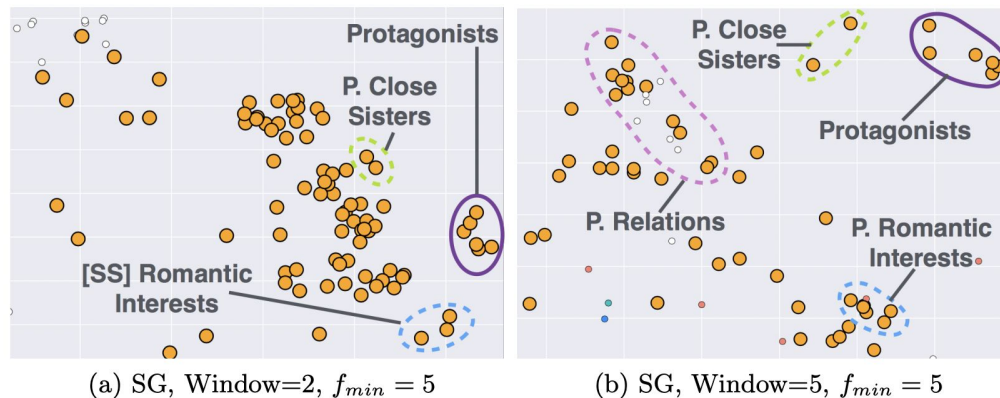


Fig. 4: Character embeddings from the Austen dataset.



Что вообще делать с векторными моделями?

Наиболее понятный способ, которым из модели извлекается информация, это получение определенного количества (конвенционально — 10) слов-векторов наиболее схожих с данным (наиболее близких к данному по косинусному расстоянию).

Этот список дает наиболее полное представление о положении вектора в векторном пространстве. В лингвистических исследованиях такие данные представляют особенности семантики слова в том корпусе, на котором строилась модель.



Как это обычно выглядит?

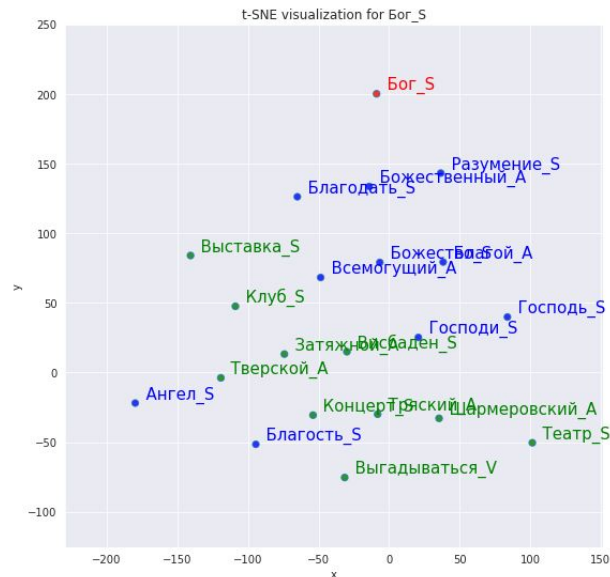
- Например, ближайшими соседями слова «дом» в модели, построенной на газетных текстах, будет «здание», «дворец», «комната», «интернат», «больница»,
- а в модели, построенной на поэтических текстах XX века, — «фундамент», «деревня», «крыльцо», «стол».

И те, и другие представляют собой законные ассоциации,

- но в первом случае актуализированы «официальные» воплощения домов как социально значимых учреждений (больница, интернат, дворец тут тоже из сочетания типа «дворец культуры»),
- а во втором — дома как образной точки в пространстве, вступающей в ассоциативные связи с символически нагруженными предметами (крыльцо, стол).

Векторная модель семантики Толстого

- Орехов Б.В. Индивидуальная семантика Л. Н. Толстого в свете векторных моделей // Terra Linguistica. 2023. Т. 14. № 4. С. 119–129. [DOI: 10.18721/JHSS.14409](https://doi.org/10.18721/JHSS.14409)
- Корпус: около 8 млн лемматизированных (mystem) слов, около 100 тыс. уникальных слов в модели только с частотностью 2+ поэтому вокабуляр модели 45 тыс. слов
- Сравнение с моделью для НКРЯ



любить для Толстого:



Толстой:

полюбить VERB
уважать VERB
ненавидеть VERB
страстно ADV
любящий ADJ
ценить VERB
презирать VERB
дорожить VERB
ближний ADJ
жалеть VERB

НКРЯ:

обожать VERB 0.74
полюбить VERB 0.70
любить NOUN 0.68
уважать VERB 0.66
нравиться VERB 0.66
ненавидеть VERB 0.64
любимый VERB 0.62
любить ADJ 0.60
боготворить VERB 0.59
презирать VERB 0.57



Обожать, боготворить: *ненастоящая любовь*

— Ничего нет ужасного, — сказал Новодворов, прислушивавшийся к разговору. — Массы всегда **обожают** только власть, — сказал он своим трещащим голосом. — Правительство властвует — они **обожают** его и ненавидят нас; завтра мы будем во власти — они будут **обожать** нас...

(Воскресение)

Девушка эта рассказала Николаю, как она с детства еще, по портретам, влюбилась в него, **боготворила** его и решила во что бы то ни стало добиться его внимания.

(Хаджи-Мурат)

Word2vec русской поэзии



Стих и проза

Орехов Б. В. Стихи и проза через призму дистрибутивной семантики // Острова любви БорФеда: Сборник к 90-летию Бориса Федоровича Егорова / ИРЛИ РАН; СПБНИИ РАН; Союз писателей Санкт-Петербурга; Ред.; сост. А. П. Дмитриев и П. С. Глушаков. — СПб.: Издательство «Росток», 2016. — С. 652–655.

Сравним модели для стихотворных и прозаических текстов:

- наиболее частотные существительные поэтического корпуса одновременно и наиболее сильно отличаются по составу квази;синонимов в стихотворных и прозаических текстах
- *земля, любовь, человек, час* не имеют общих квази;синонимов



Семантика поэтизмов

- Самые частотные существительные поэтического корпуса: *сердце, душа, жизнь* (более 20 тыс. вхождений), каждое имеет не более двух пересечений в списке квази-синонимов с существительными прозаической модели.
- Для лексемы «сердце» такими пересечениями являются квази;синонимы душа и грудь, а для «души» — сердце и дух. Единственным общим квази;синонимом для прозаического и стихотворного употребления слова «жизнь» является «житие».
- семантика слова в стихотворном контексте серьезно отличается от той, которая актуализируется в контексте прозаическом.

Word2vec русской прозы



Разметка

Character-textid-2765660281819920316-charid-752 в школе учился хорошо, благодаря **Character-textid-2765660281819920316-charid-752** хорошим способностям, но был ленив и шалун и потому вышел из последних; но, несмотря на **Character-textid-2765660281819920316-charid-752** всегда разгульную жизнь, небольшие чины и нестарые годы, **Character-textid-2765660281819920316-charid-752** занимал почетное и с хорошим жалованьем место **Character-textid-2765660281819920316-charid-754**.

Место это **Character-textid-2765660281819920316-charid-752** получил чрез **Character-textid-2765660281819920316-charid-757**, **Character-textid-2765660281819920316-charid-211**, занимавшего одно из важнейших мест в министерстве, к которому принадлежало присутствие; но если бы **Character-textid-2765660281819920316-charid-211** не назначил **Character-textid-2765660281819920316-charid-752** на это место, то чрез **Character-textid-2765660281819920316-charid-760**, **Character-textid-2765660281819920316-charid-761**, **Character-textid-2765660281819920316-charid-762**, **Character-textid-2765660281819920316-charid-763**, **Character-textid-2765660281819920316-charid-764**, **Character-textid-2765660281819920316-charid-765**, **Character-textid-2765660281819920316-charid-766**, **Character-textid-2765660281819920316-charid-767** получил бы это место или другое подобное, тысяч в шесть жалованья, которые **Character-textid-2765660281819920316-charid-752** были нужны, так как дела **Character-textid-2765660281819920316-charid-752**, несмотря на достаточное состояние **Character-textid-2765660281819920316-charid-524**, были расстроены.



Две модели

Мы построили две word2vec-модели на корпусе из 450 текстов, Continuous Bag of Words, контекстное окно 10 слов:

одна учитывала контексты и полнозначные слова в лемматизированном виде, встретившиеся в корпусе более 1 раза

другая учитывала исходные словоформы, включая служебные, но также с порогом частотности 2.

В обеих моделях 1208 векторов персонажей, первая содержит 547986 слов,

вторая 732318 слов.

Вторая модель должна более детально передавать авторскую манеру портретирования = стилистические особенности грамматики.



Опубликованные модели

```
@misc {dh_cloud_2024,  
  author = { {DH CLOUD} },  
  title = {  
w2v-russian-19c-fiction-lemmas  
(Revision 2d3e2cc) },  
  year = 2024,  
  url = {  
https://huggingface.co/dhcloud/w2v-russ  
ian-19c-fiction-lemmas },  
  doi = { 10.57967/hf/2437  
},  
  publisher = { Hugging Face }  
}
```

```
@misc {dh_cloud_2024,  
  author = { {DH CLOUD} },  
  title = {  
w2v-russian-19c-fiction-forms (Revision  
dd6ee8b) },  
  year = 2024,  
  url = {  
https://huggingface.co/dhcloud/w2v-russ  
ian-19c-fiction-forms },  
  doi = { 10.57967/hf/2436  
},  
  publisher = { Hugging Face }  
}
```



Персонажи



Анна Каренина (формы)

Наиболее близкими персонажами к Анне Карениной в модели, построенной на **словоформах**, оказались **княжна Марья** из «Войны и мира» Толстого, **графиня** из повести Н. А. Бестужева «Русский в Париже 1814 года», **Мери** из повести Е. Ковалевского «Петербург днем и ночью», **Зинаида Николаевна** из романа «В места не столь отдаленные» К. Станюковича.

Это женские персонажи, которые объединены **способом их портретирования** как персон, испытывающих искренние чувства по контрасту с окружающей их социальной средой.



Анна Каренина (леммы)

Анна оказывается похожа сразу на нескольких женских персонажей Всеволода Соловьева: **Груню, Софью Сергеевну** из «Последних Горбатовых» (1886) и Екатерину из «Волхвов» (1889).

Это не значит, что какой-то из этих списков более правильный.

Каренина больше похожа на княжну Марью, если иметь в виду **формы** употребленных слов, и на Груню, если иметь в виду более **абстрактные способы** создания портрета.



Вс. Соловьев и Толстой

Влияние поэтики **Толстого** на Вс. Соловьева.

Соловьев выделял роман «Анна Каренина» из потока публицистики (Рус. пис. Т. 5. С. 742), хотя биографически и идейно в большей степени был близок Достоевскому.

В то же время: «Фатальная приверженность к недостойному предмету страсти, всевластие любви как “наваждение”» (Рус. пис. Т. 5. С. 743).



Платон Каратаев (леммы)

Державин | Державина | Державиным | Державину | сварливый и завистливый старик
Державин | старика Державина | Старик Державин | спящего Державина | старик
Державин | какого-то Державина [mordovezv_d_dvenadzatyj_god_1879](#) 0.6337

Вольтер | Вольтера | Вольтере | Вольтером | вашего Вольтера
[mordovezv_d_dvenadzatyj_god_1879](#) 0.6156

Мерзляков | Мерзлякову | Мерзляковым | Мерзляковы | Алексей Федорович | Алексей
Федорович Мерзляков | профессор Мерзляков | почтеннейший Алексей Федорович |
Мерзлякова | Мерзлякову ни с того ни с сего [mordovezv_d_dvenadzatyj_god_1879](#)
0.5906



Обломов (леммы)

Штольц | Обломов | Обломова | Обломову | Обломовым | Обломове | мсье Обломов | Штольце | XI Обломов | VIII Штольц [goncharov_i_oblomov_1859](#) 0.7336

Захара | Захару | своего Захара | Увещевая Захара | Захара насилу [goncharov_i_oblomov_1859](#) 0.6866

Захар | Захар Трофимыч | Захаром | Пришел Захар | VIII Захар | неотвязчивый Захар | Захар , показывая какие-то две подошвы вместо рук | Трофимыч | Захар , понявший только последние слова | Захар в отчаянии [goncharov_i_oblomov_1859](#) 0.6725

Агафья Матвеевна | Агафьи Матвеевны | Агафью Матвеевну | Агафьей Матвеевной | Добрая Агафья Матвеевна | Агафья Матвеевна с детьми [goncharov_i_oblomov_1859](#) 0.6534

Ольга | Ольги | Ольге | Ольга Сергеевна | Ольгу | Ольгу Сергеевну | Бедная Ольга | ненаглядную Ольгу | милая Ольга | кроткая Ольга [goncharov_i_oblomov_1859](#) 0.6377

Александр | Александра | Александр Федорыч | Александрю | Александра Федорыча | Александром | Саша | Александре | Александрю Федорычу | прежним Александром

[goncharov_i_obyknovennaya_istoriya_1847](#) 0.6062

Райский | Райского | Райскому | Борис Павлович | Борис | Райским | Борису | Борису Павловичу | Бориса Павловича | XVIII Райский [goncharov_i_obryv_1869](#) 0.5680



Базаров (леммы)

Павел Петрович | Павла Петровича | Павлу Петровичу | Павлом Петровичем | Павел | Кирсанов | Кирсанова | господин Кирсанов | изумленный Павел Петрович | один Павел Петрович [turgenev_i_otzy_i_deti_1862](#) 0.6954

Анна Сергеевна | Одинцова | Анны Сергеевны | Анну Сергеевну | Анне Сергеевне | Одинцову | Анна Сергеевна Одинцова | Одинцов | покойный Одинцов | расстановкой Одинцова [turgenev_i_otzy_i_deti_1862](#) 0.6822

Миропа Дмитриевна | Аггей Никитич | Миропе Дмитриевне | Миропу Дмитриевну | Аггея Никитича | Аггея Никитича Миропа Дмитриевна | Жила Миропа Дмитриевна | Рыжовых громадного капитана Аггея Никитича [pisemskij_a_masonry_1880](#) 0.6790
Николай Петрович | Николая Петровича | Николаю Петровичу | Николай | Николая | Николая - чудотворца | бедному Николаю Петровичу | Николаем Петровичем [turgenev_i_otzy_i_deti_1862](#) 0.6720

Борис Андреич | Бориса Андреича | Борису Андреичу | Борисе Андреиче | Борис Андреич, который, как видно, ожидал этого приглашения с некоторым нетерпением | неугомонный Борис Андреич [turgenev.dva_priyatelya](#) 0.6555

Дмитрий | Дмитрия | Дмитрию | Дмитрием | чудесным Митей | Один Дмитрий | мать Дмитрия | моего Дмитрия | Дмитрий, старавшийся понимать любовь | Дмитрием, который, расхаживая взад и вперед, поправлял шейей галстук [tolstoy.yunost](#) 0.6510

Нежданов | Сипягин | Сипягина | Нежданова | Нежданову | Неждановым | Сипягиным | Сипягиной | Сипягину | Сипягиных [turgenev_i_nov_1877](#) 0.6489

Катя | Аркадий | Катю | Катей | Аркадия | Кате | Аркадию | Аркадием | Кати | Катей Аркадий [turgenev_i_otzy_i_deti_1862](#) 0.6461

Петр Васильич | Петра Васильича | Петру Васильичу | Петром Васильичем | бедного Петра Васильича | бедный Петр Васильич | изумленный Петр Васильич [turgenev.dva_priyatelya](#) 0.6374



Пьер Безухов (формы)

князь Андрей | Князь Андрей | князя Андрея | князю Андрею | Андрей | князем Андреем | Князю Андрею | князе Андрее | Андрея | Князя Андрея [tolstoj_l_vojna_i_mir_1868](#) 0.8655

Ростов | Николай | Ростова | Ростову | Ростовых | Николая | Ростовым | Николаю | Николаем | Ростовы [tolstoj_l_vojna_i_mir_1868](#) 0.8359

Пьер | Пьера | Анна Павловна | Анны Павловны | Пьером | Анне Павловне | мсье Пьер | Анну Павловну | Пьера Анна Павловна | сам Пьер [tolstoj_l_vojna_i_mir_1868](#) 0.7687

Левин | Степан Аркадьич | Левина | Левину | Левиным | Левине | Левиных | Степан Аркадьич Левину | Степан Аркадьич, который любил физиологию | охотник Степан Аркадьич [tolstoj_l_anna_karenina_1877](#) 0.7645

Алексей Александрович | Алексея Александровича | Алексею Александровичу | Каренина | Алексеем Александровичем | Каренин | Каренину | Алексей | Карениным | Алексее Александровиче [tolstoj_l_anna_karenina_1877](#) 0.6914

Вронский | Анна | Вронского | Вронским | Вронскому | Анны | Анне | Анну | Вронском | Аннушка [tolstoj_l_anna_karenina_1877](#) 0.6839

Пьеру | Пьеру духовною [tolstoj_l_vojna_i_mir_1868](#) 0.6777

Глинский | Глинского | Глинскому | Глинским | г. Глинский | Глинском | Г. Глинский | бедный Глинский | самого Глинского | г. Глинскому [bestuzhev_n_russkij_v_parizhe_goda_1860](#) 0.6717

Вронский | Вронского | Вронскому | Алексей | Вронским | Алексей Вронский | Алексею | Алексея Вронского | Алексеем | Вронском [tolstoj_l_anna_karenina_1877](#) 0.6684



https://t.me/universitates_podcast